HOWARD UNIVERSITY

Investigating an Automated Method to Explore Mesoscale Convective Complexes in West Africa

A Dissertation
Submitted to the Faculty of the
Graduate School

of

HOWARD UNIVERSITY

in partial fulfillment of
the requirements for the
degree of

DOCTOR OF PHILOSOPHY

Program of Atmospheric Sciences

by

Kim Dionne Whitehall

Washington, D.C.
May 2014

HOWARD UNIVERSITY
GRADUATE SCHOOL
PROGRAM OF ATMOSPHERIC SCIENCES

DISSERTATION COMMITTEE

Belay Demoz, Ph.D.
Chairperson

Gregory Jenkins, Ph.D.

Chris Mattmann, Ph.D.

Mugizi Rwebangira, Ph.D.

Claire Monteleoni, Ph.D.
Assistant Professor
School of Engineering and Applied Science
Department of Computer Science
George Washington University
Washington, D.C.

Gregory Jenkins, Ph.D.
Dissertation Advisor

Candidate: Kim Whitehall
Date of Defense: April 14, 2014

ACKNOWLEDGEMENTS

First and foremost, I wish to give thanks to the Almighty Father. I am so thankful for the village of supporters God has blessed me with throughout my life, and this process.

I wish to sincerely thank my advisor, Dr. Gregory Jenkins, for taking me on as a doctoral student during my first semester at Howard University. Greg's support and guidance throughout my graduate school career, and belief in the research and my abilities to complete it will never be forgotten.

I would also like to thank the members of my dissertation committee – Dr. Belay Demoz, Dr. Robert Rwebangira, Dr. Chris Mattmann and Dr. Claire Monteleoni. I appreciate the time you took to understand my research and the eagerness with which you encouraged it. I especially would like to thank Chris who I am forever indebted to for his tutelage, encouragement, and mentorship even far beyond the academics.

The scope of this research lead to interesting collaborations and discussions with the Regional Climate Modeling Evaluation System (RCMES) team at the NASA Jet Propulsion Laboratory. As such, I wish to thank the entire team, and make special mention of Cameron Goodale and Paul Zimdars, for their contributions along the way that improved the quality and content of this dissertation research.

I would also like to acknowledge some of the sources of funding that supported the research for this degree, which include primary funding from the National Science Foundation (NSF) grant AGS-1013179, and secondary funding from NSF awards PLR-1348450, ICER-1343800, ACS-1125798, and GEO-1229036 and GEO-1343583. Efforts were also partially supported by the Jet Propulsion Laboratory, managed by the California Institute of Technology under a contract with the National Aeronautics and Space Administration.

I must also acknowledge and thank the administrative staffs of the International Student Services Office, the Graduate School, the Howard Program in Atmospheric Sciences, and the Physics Department for their guidance throughout the three years.

Finally, I wish to sincerely thank my village of family and friends, and acknowledge them for their support, friendships, and encouragement through it all - my mother who everyday offered words of wisdom and encouragement; Marquez, my significant other, for his patience, support, and invaluable ability to lightening my mood and aid with refocusing; Alexander, my darling nephew, whose smile brightened the worse of days; and my friends for their tireless support.

DEDICATION

To my loving, supportive mother, who never ceased believing this was possible. To my caring, encouraging and most patient love, Marquez, whose personal sacrifices will never be forgotten. To my dearest friend, Barbara, who started this journey with me, and watched it come to completion from Heaven.

ABSTRACT

Mesoscale convective complexes are convectively driven, high impact weather systems with durations of approximately 10-12 hours, and are large contributors to daily and monthly rainfall totals. In West Africa, approximately 40 mesoscale convective complexes contribute an estimated one-quarter of the total rainfall amounts between July and September annually. As such, an understanding of the lifecycle, characteristics, frequency, and seasonality of these weather features is critical for climate studies, agricultural and hydrological studies, and for disaster management. Identification criteria of mesoscale convective complexes exist for infrared satellite data, but the spatial extent and the spatio-temporal variability of the convective characteristics of these mesoscale convective complexes make rainfall characterization difficult, even in dense networks of radars and / or surface gauges. Hence, fully automated methods are required to explore mesoscale convective complexes in long-term infrared satellite data, and to determine their characteristics from other datasets, such as precipitation rate satellite datasets.

Automated identification methods of mesoscale convective complexes are based on forward- and / or backward-in-time spatial-temporal analyses of infrared satellite data, and usually incorporate a manual component to verify the features and / or characterize the associated precipitation. These existing identification and precipitation characterization methods are not readily transferable to "big data" such as satellite-derived datasets, thus hindering comprehensive studies of these features, both at weather and climate timescales. In recognizing these limitations and the growing volume of satellite data, this study explores the applicability of graph theory to creating a fully automated method for identifying mesoscale convective systems in satellite datasets. The framework for such a method is provided in this work.

The results indicate that applying graph theory innately handles the complexity of the

mesoscale convective complexes, thus eliminating a manual verification stage. Furthermore, implementing a graph theory based method identifies individual and embedded features in infrared satellite datasets and extracts the associated precipitation rate data from other datasets for comparison in a seamless fully automated manner. The results also establish that a graph theory based method allows for studies over periods and spatial domains longer and larger than individual events.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AEJ	African Easterly Jet
AEW	African Easterly Wave
AMMA	African Monsoon Multidisciplinary Analysis
ASF	Apache Software Foundation
BFS	Breath first search
CC	Cloud Cluster
CE	Cloud Element
CERES	Cloud and Earth Radiant Energy Sensor
CORDEX	Coordinated Regional Climate Downscaling Experiment
CPC	Climate Prediction Center
CRU	Climate Research Unit
DFID	Depth first iterative deepening
DFS	Depth first search
ENSO	El Niño Southern Oscillation
ESGF	National Oceanic and Atmospheric Administration
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
GARP	Global Atmospheric Research Program
GATE	GARP Atlantic Tropical Experiment
GCM	Global Climate Model
GMS	Geostationary Meteorological Satellite
GTG	“Grab ‘em, Tag ‘em, Graph ‘em”

IDL	Interactive Data Language
IPCC	Intergovernmental Panel on Climate Change
IR	Infrared
ISCCP	International Satellite Cloud Climatology Project
ITCZ	Inter-Tropical Convergence Zone
ITF	Inter-Tropical Front
JIFRESE	Joint Institute for Regional Earth System Science and Engineering
JPL	NASA's Jet Propulsion Laboratory, California Institute of Technology
LIFO	Last in, First out
LIS	Lighning Imaging Sensor
MASCOTTE	Maximum Spatial Correlation Tracking Technique
MCC	Mesoscale convective complex
MCS	Mesoscale convective system
MW	Microwave
NASA	National Aeronautics and Space Administration
NCEP	National Centers for Environmental Protection
NETCDF	Network Common Data Form
NOAA	Earth System Grid Federation
NWS	National Weather Service
OCW	Apache Open Climate Workbench
PR	Precipitation Radar
RCM	Regional Climate Model

RCMED	Regional Climate Model Evaluation Database
RCMES	Regional Climate Model Evaluation System
RCMET	Regional Climate Model Evaluation Toolkit
SEVIRI	Spinning Enhanced Visible Infrared Imager
SST	Sea-Surface Temperature
T_B	Brightness temperature
TEJ	Tropical Easterly Jet
TMI	TRMM Microwave imager
TOOCAN	Tracking of Organized Convection
TRMM	Tropical Rainfall Measuring Mission
VIRS	Visible and infrared scanner
VIS	visible
VM	Virtual Machine
WAM	West African Monsoon
WCRP	World Climate Research Programme

CHAPTER 1. INTRODUCTION

1.1 Research Problem and Broader Impacts

Rainfall variability in current and future climates is a major concern for countries planning adaptation and mitigation strategies. For example, for impact studies on hydrological processes or agriculture, it is imperative to study the spatial and temporal scales of the processes influencing the partition of rainwater into the various aspects of the hydrological cycle. Because of this, large-scale predictions are not ideal. The World Climate Research Programme (WCRP) devised the Coordinated Regional Climate Downscaling Experiment (CORDEX) to address climate issues on regional scales, noting that certain regions, such as Africa, are especially vulnerable to climate variability and climate change. This is mostly due to the limited scientific understanding of atmospheric processes in those regions, the lack of adequate financial resources and technological resources, and limited human capacity to adequately adapt or mitigate (IPCC, 2007; Giorgi, Jones & Asrar, 2009).

The region of West Africa has a population of 245 million persons, with 65 percent of those persons living in rural areas (West Africa, n.d.). The average annual income is 309 US dollars, approximately 161 US dollars lower than that in the Sub-Saharan area. The major economic activity in most West African countries is agriculture, which is closely coupled to the weather and climate. Inaccurate timing of seasons (wet / dry) or ill-forecasted accumulations can lead to devastating effects in both arable and pastoral agriculture.

The convective nature of tropical rainfall, the limited data available, the limited scientific understanding of the interactions between known drivers that generate rain, and the various spatial and temporal scales at which the major drivers respond are amongst the problems with forecasting rainfall in West Africa (Waliser *et al.*, 2012). Global climate models (GCMs) and

regional climate models (RCMs) can resolve large-scale (synoptic) features such as the Inter-tropical Convergence Zone (ITCZ), and represent convective precipitation in their parameterizations. However, there is high uncertainty in models' ability to represent rainfall as it is a parameterized process, and thus, the lack of the scientific understanding of the processes in the current climate, limited data and limited computer resources restrict the complexities of the parameterizations. For example, Nukulin *et al.* (2012) examined the accuracy of rainfall representation within an ensemble of CORDEX-Africa RCMs against two types of satellite derived data and (coarse) rain gauge data, and determined there were issues with the timing and the positioning of rainfall events within the ensemble, especially concerning diurnal rainfall peaks. To effectively model rainfall, knowledge and representation of the diurnal response of the convection within the climate model would be necessary (D'Amato & Lebel, 1998; Slingo, 1987). Additionally, coupling of the diurnal cycle and the large-scale climate dynamics is also necessary (Duvel, 1989).

The distribution of the annual rainfall in West Africa is minimum (~50mm) towards the Sahara Desert and maximum along the coastal regions close to the equator (~4000mm in some areas). The distance between the northern part of West Africa to the southern part is ~ 2000 km indicating a large change in rainfall totals across approximately 20 degrees of latitude. Most of the rain within the region falls in the afternoon and at night. The rainfall is strongly correlated with mesoscale convective systems (MCSs), is highly spatially and temporally variable, and is heavily influenced by land-surface properties and large-scale circulations. MCSs are well-organized areas of thunderstorms that persist for several hours (Ray, 1986). A subclass of MCSs is mesoscale convective complexes (MCCs). MCCs are large convective weather systems that are associated with large precipitation totals in short durations. According to Maddox (1980b),

the identification of MCCs utilizes infrared satellite imagery. This method is preferable, as the spatial expanse of MCCs and large-scale variability in their convective characteristics make them difficult to characterize, even in dense networks of radars and / or surface rain gauges. MCCs are most common along the western coastline countries of West Africa, and within the Sahel region (Laing & Fritsch, 1993). Furthermore, as noted by Taylor, Lambin, Stephenne, Harding and Essery (2002) and Mathon, Laurent and Lebel (2002), surface observations in West Africa are few and / or unreliable with respect to duration and frequency of measurements, making it is beneficial to observe the rainfall characteristics via satellite images. However, satellite images are limited in their accuracy and techniques to derive variables from satellite sensors are a continuous source of research, as noted by Arkin (1979), Stephens (1990), Nalli and Stowe (2002) and Nalli and Reynolds (2006).

Predictability of rainfall should consider all scales that it occurs on (Hoskins 2013). As such, it is expected that MCCs should be represented in climate model outputs. However, their representation in this data set would not follow the identification criterion used with meteorological analysis, such as criterion developed by Maddox (1980b) because the convective, microphysics and land-surface processes that produce MCCs occur on the sub-grid scale in the models. Additionally, climate models do not necessarily forecast the variables used in MCC identification criterion on the scales required. Instead, representation of MCCs within existing climate models intuitively would be embedded in the monthly and seasonal rainfall predicted. However, this lends to the concern of the representation of MCCs in the regional climate model's (RCM) simulated monthly totals, and if the MCCs are being represented as extremes or areas of high precipitation. Another concern is if RCMs are accurately reproducing MCC rainfall totals and variability in current climates and what methods can be used to evaluate the MCC rainfall

variability in current observation datasets of rainfall. In order to address these concerns, a means of determining the characteristics of MCCs (and other mesoscale convective systems) in variables that are commonly measured, such as precipitation, is necessary. Currently, no such method exists.

This research proposes to explore a transferable, fully automated method for identifying mesoscale convective systems, specifically mesoscale convective complexes, over long records of satellite datasets representing different variables. In this research the variables considered are brightness temperature and precipitation rate, though the method is extensible to other variables e.g. precipitation totals. It is recognized that the final product of such a method would engender a new scope of evaluations of MCC features in various observation and model datasets, but, only the exploration stages of the concept of such a method are presented in this work. It is further proposed to utilize functionality from an Apache Software Foundation (ASF) project. The ASF is an open-sourced community-led software development platform that provides software products for the good of the public. Specifically in this research the ASF Top Level Project, Apache Open Climate Workbench[1] (OCW) will be utilized for data extraction, manipulation such as regridding, metrics calculations, and visualization for handling the various datasets in the method. Leveraging Apache OCW will not only efficiently handle the "big data" issues inherent within this problem, but will also encourage a unified information technology approach (Mattmann, 2013).

[1] http://climate.apache.org/

1.2 Research Objectives

The preceding section highlighted, on the high-level, the need for an application that address data methods for analyzing satellite data so that comprehensive studies related to MCCs can be conducted. Specifically, the following points have been identified as a need:

(1) An automated method to allow for comparisons of the MCCs (and other large MCSs) identified in the infrared satellite datasets with other datasets e.g. precipitation data from various sources e.g. precipitation rate satellite data or rain gauge data, to determine precipitation characteristics of the features and an appropriate database. This method should seamlessly handle datasets of various spatial and temporal resolutions.

(2) An automated method to identify weather-scaled phenomenon, large-scaled MCSs, in long-term high-resolution infrared satellite datasets, to produce datasets related to the weather feature. This method should seamlessly handle the big data issues – voluminous, variety and velocity – associated with dealing with satellite data over long-term periods, and climate model data

This research will examine a method that compliments a framework for a fully automated tool for identifying mesoscale convective complexes in West Africa.

Objective 1: Identify or create a method that can identify mesoscale convective complexes in infrared and precipitation satellite datasets without manual intervention.

Objective 2: Explore the applicability of that method to addressing the identification of mesoscale convective complexes over West Africa in satellite datasets durations longer than a single event.

Objective 3: Create the method in an open-sourced environment (Mattmann *et al.*, 2012). This encourages reuse of existing methods for aspects of the application and promotes overall contributions (development) of the code associated with the application long after the scope of this project. Additionally, by creating the tool in an open-sourced environment, the software development must follow best software development practices such as using a component / module based architecture and iterative development with verification of the quality, thus making it a viable application in other environments.

1.3 Roadmap to the Document

Background information regarding rainfall variability in the region of interest, West Africa will be provided in Chapter 2. Specifically, the significance of rainfall variability in West Africa and a literature review of its drivers will be presented. The data sets and tools used within this study will then be briefly discussed in Chapter 3. The Apache OCW will also be explained in that chapter. Chapter 4 will present the new automated method for identifying large-scaled weather systems, specifically MCCs, in infrared and precipitation satellite datasets. Chapter 5 will present case studies demonstrating the accuracy of the automated method in different scenarios. Finally, Chapter 6 will present the conclusions of the research and future directions for the work.

CHAPTER 2. BACKGROUND OF RAINFALL VARIABILITY IN WEST AFRICA

The rainfall variability in West Africa has local and international socio-economic and climatic impacts (Redelsperger *et al.*, 2006). Within the scope of this project, it is important to understand the not only the characteristics of the weather feature that will be the focus of this research, but also the influences on its formation. This chapter provides a summary of rainfall variability in West Africa and outlines limitations with rainfall climate projections in this region.

2.1 A Summary of Rainfall Variability in West Africa

2.1.1 Rainfall Distribution in West Africa

West Africa is considered to be one of the semi-arid regions of the world, and inherently an area of unreliable rainfall from year to year. The seasons in West Africa are defined as wet or dry, where the dry season extends from the end of November until April, and the wet or rainy season extends from May until October. The rainy season is bimodal in pattern where the first period extends from the end of May until August, and the second, shorter period from September to October. The maximum rainfall occurs within the first rainy period, around late August. In general, the rainfall in West Africa is highly spatially and temporally variable. Observations indicate that there are distinct zones of homogenous rainfall following lines of latitude within the region (Nicholson, 1980).

Within the northern part of West Africa, the annual rainfall is ~50 mm. The rainfall totals increase southward towards the equator, with maximum areas of rainfall along the equatorial coastal areas totaling as much as 4,000 mm annually (Figure 2.1). The homogenous rainfall zones are broken up within the region referred to as the Sahel. In most of West Africa,

observations indicated the rainfall totals of the 1970s and 1980s were ~180mm – 200mm less as compared to those of the 1950s and 1960s (Le Barbé, Lebel & Tapsoba, 2002).

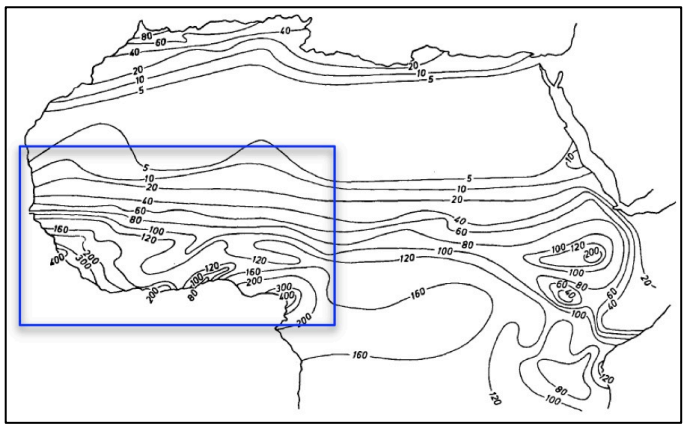


Figure 2.1. Annual rainfall (cm) in West Africa, taken from Nicholson (1980). The blue box provides an estimate of the area referred to as West Africa within this study.

Events such as the famines in the Sahel in 1972 – 1974 and 1983 – 1985 are so distinct, that they not only aroused international interests to the food security and famine issues in these areas, but also sparked the interest of atmospheric scientists to investigate the mechanisms responsible for the anomalous events and possibly their predictability. Observations also indicated that annual rainfall totals exceeding the annual average have been occurring more frequently in the 2000s.

2.1.2 Drivers of Rainfall Variability in West Africa

The climate system is commonly defined as all air, water, ice, minerals, rocks and living organisms on Earth. Within the climate system, transfers of energy, momentum and mass are simultaneously occurring at several temporal and spatial scales (Ray, 1986). Weather is

generated by instabilities created in variables such as pressure, temperature, humidity, as a result of changes in the climate system due to the response of a variable on a given scale. Within the atmosphere, temporal scales that result in changes of variables range from short term (in the order of microseconds) for turbulent and frictional influences, to long term (in the order of years and centuries) for planetary scale circulations within the system. Figure 2.2 provides an illustration of common weather phenomenon in the Tropics and their spatial and temporal scales.

For the purpose of this study, attention is given to the mesoscale weather phenomenon. Mesoscale weather features have a temporal scale ranging from hours to one day, and a spatial resolution ranging from 10^2 m to 10^6 m. Continuity exists between the scales, for example, mesoscale weather features that do not decay, continue to grow to meet the criteria of synoptic features.

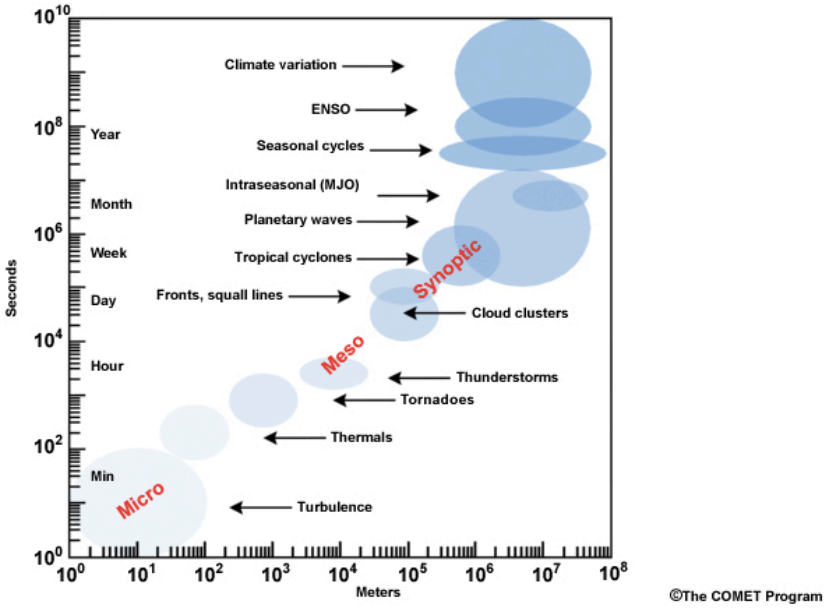


Figure 2.2. Classification of meteorological scales for typical weather phenomena in the Tropics (Laing and Evans, 2011).

Climate is the statistical description regarding the average and variability of weather and other climate system parameters over a given location, such as a country, a region, or a planet, and over a given period, such as months, thousands of years, or millions of years (Frequently Asked Questions, n.d.). Climate variability refers to variations in the mean state, standard deviation and other statistics of climate on temporal scales that exceed any one weather event. According to the IPCC (2007) seasonal variability is used to define the changes of a phenomenon (weather variable or feature) within a 12-month period. Intraseasonal variability refers to the changes within a season, and interannual variability is used to define the changes between years. The term climate change is used to define statistical significant and lasting change in the mean state and / or variability of the climate that persisted for extended periods. The reason for the climate change may be linked to natural processes (internal or external to the climate system) or to anthropogenic changes in the climate system. Like weather, climate classification also has a spatial component. In general, a direct correlation exists such that a short temporal scale correlates with a small spatial scale, e.g. seasonal variability studies focus over geographical regions such as countries and cities, and vice versa.

The interannual variability of the rainfall in the Sahel region, and generally West Africa, has been linked to large-scale tropical convection associated with global and regional climate features as in Figure 2.3. From observations during Global Atmospheric Research Program (GARP) Atlantic Tropical Experiment (GATE), which commenced in the summer of 1974, identified and confirmed a relationship between the diurnal behavior of convection and rainfall amounts indicating an intraseasonal scale of rainfall variability also exists (Kuettner, 1974; Murakami, 1979). The interseasonal variability of rainfall has been linked to regional drivers, as well as smaller scale weather features such as mesoscale convective systems (MCSs).

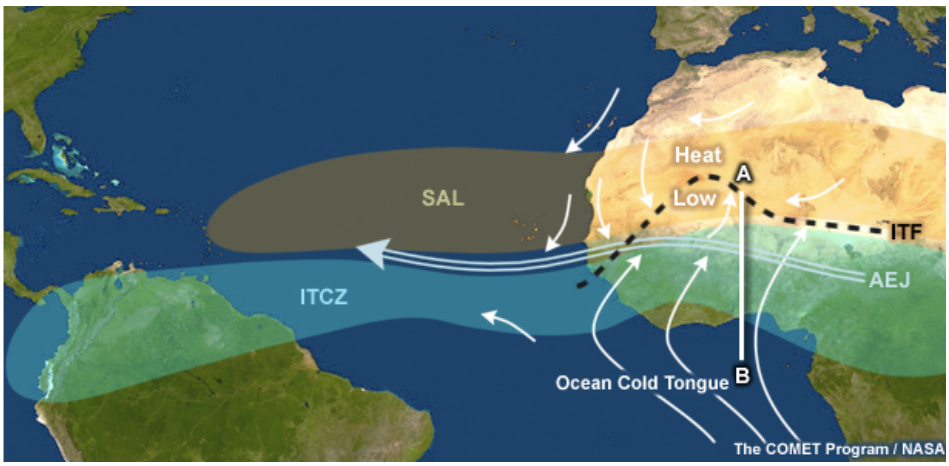


Figure 2.3. Large-scale and regional features driving the West African rainfall variability (Laing and Evans, 2011).

2.1.3 Large-scaled Circulations

The Hadley circulation is a thermally-driven meridional circuit comprising of warm equatorial air rising and flowing toward the poles in the upper atmosphere, cooling while being transported aloft, and then sinking to the surface at subtropical latitude (Hartmann, 1994; Hastenrath, 1991; Holton, 2004). The Hadley circulation varies both temporally and spatially. These variations are closely coupled to the meridional solar heating and absorption due to seasons, according to the Earth-Sun relationship. During the boreal winter, there is cross-equatorial overturning within the Northern Hemisphere Hadley circulation i.e. the rising branch is in the Southern Hemisphere. The rising branch of the Hadley cell is associated with wet climates where precipitation is heavy and exceeds the evapotranspiration for much of the year. The sinking branch is associated with desert climates.

The Inter-tropical Convergence Zone (ITCZ) is defined as the zone of the ascending branches of the Northern and Southern hemisphere Hadley circulations and is driven by the large-scale continuity of mass and momentum (Holton, 2004). The vertical thermodynamic

structure of the ITCZ indicates that the atmosphere is conditionally stable above the middle troposphere, and as such, the large-scale upward mass in this environment would lead to cooling aloft (and thus would not satisfy the heat energy budget in the tropics). In order for this energy to be transported aloft into the upper troposphere, the equivalent potential temperature has to be conserved as a parcel is traveling upwards. This is possible through deep cumulus cloud formations such as cumulonimbus clouds. Hence, the ITCZ is actually observed (especially visually on satellite images) as clusters of vigorous cumulonimbus and convection clouds divided by clear patches along the convergence zone of the Hadley cells. In West Africa, the north-south movement of the ITCZ as the seasons change between hemispheres is associated with the onset and ending of the rainy season.

2.1.4 Regional Circulations

The Tropical Easterly Jet (TEJ) is an upper tropospheric circulation (observed between 200 mb and 100 mb) over Asia / Indian Ocean and West Africa during the boreal summer, where winds in the jet stream are ~40 ms^{-1}. The Hadley circulation caps the TEJ. It is noteworthy that though there are two components of the TEJ, the information provided within refers to the branch over West Africa. Because the TEJ results from a thermal wind circulation, its location is heavily influenced by the location of the African continent and the Saharan Desert, where the mid-tropospheric thickness is large and the gradient reversed due to the intense heating over the desert. It is persistent around 10 °N, though over more continental regions it can be higher, tending to 15 °N (Fontaine & Janicot, 1992). Rainfall maximum totals are found on the southern side of the jet, and rainfall minimum totals to the northern side (Koteswaram, 1958). Excessive rainfall in a given year in West Africa is associated with an abnormally fast TEJ and vice versa (Fontaine & Janicot, 1992). Further analysis of dry and wet days within the rainy season by El

Houssein and Decleir (1998) showed that rainy days are always correlated with a significant increase of the speed of the TEJ, and dry days do not have this distinct characteristic. This small temporal scale finding indicates that the rainfall variability observed may be more keenly dependent on daily changes than regional changes.

The African Easterly Jet (AEJ) also referred to as the West African Jet, is a mid-level tropospheric jet stream observed from April to November that forms as a result of the warming of the Sahara Desert and the surface high-pressure system associated with that heating i.e. the Saharan High (Koteswaram, 1958). The Saharan High leads to strong positive meridional geopotential gradients between the desert and the oceanic air to the south of the region, causing a strong geostrophic easterly flow (thermal wind) of ~10 ms^{-1} at 650 mb and ~13 °N, although it is more northward in July and August (Burpee, 1972). As illustrated in Figure 2.4, Cook (1999) demonstrated that the formation of the AEJ requires summer insolation (sensible heat fluxes) and dryness (latent heat fluxes) over Saharan Africa. Thorncroft and Blackburn (1999) showed that the AEJ (and the ITCZ) contribute to the unstable environment in the rainy season by reversing the potential vorticity gradient throughout the atmosphere.

In general, the location of the jet is important, as a strong African Easterly Jet (AEJ) is associated with low rainfall totals over West Africa. Furthermore, the largest wind anomalies occur at 10 °N and 5 °N (Duvel, 1989). The AEJ is associated with African Easterly Waves (AEWs) formation – disturbances generated by baroclinic-barotrophic instability within the jet (Burpee, 1972). The combination of the Tropical Easterly Jet (TEJ) and the AEJ provide a zonal circulation of mass, moisture and energy within the region in a similar sense to the global Walker circulation.

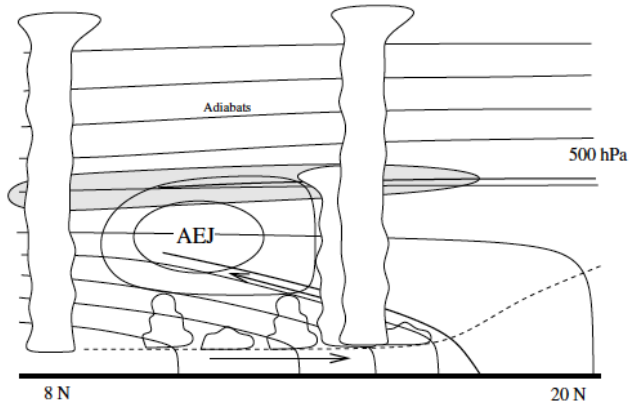


Figure 2.4. A schematic of the African Easterly Jet (Parker, Thorncroft, Burton, & Diongue-Niang, 2005).

Meteorologists and researchers alike have identified a boundary between dry and hot airmass to the north of West Africa, and wet and cooler airmass to the south. This boundary has been dubbed the Inter-Tropical Front (ITF) and is properly defined by two criteria at the surface: (1) a pressure minimum and a boundary between northerly and southerly flow; and / or (2) dew-point temperatures of roughly 15.5 °C (Beer, Greenhut & Tandoh, 1977). The geopotential height minimum is considered within the criteria for location above the surface, instead of the pressure minimum.

2.1.5 Sea-Surface Temperature Anomalies

Warm sea surface temperatures (SSTs) imply low atmospheric pressure at the surface that lead to rising air and an area favorable to cloud development. In addition to these heat fluxes, ocean basins also provide moisture transfers into the atmosphere. SSTs have been linked to rainfall through the enhancement of surface pressure gradients between the oceans and the land

that influences surface wind flow, and leads to convergence and cloud formation. A common example of such an interaction on the local / mesoscale scale is the sea-breeze phenomenon (Hastenrath, 1991). On the larger temporal and spatial scales, SST anomalies in the tropical ocean basins influence the location of the kinematic axis between Northern and Southern hemispheric trade winds and the near-equatorial convergence zones, and thus the location of maxima precipitation and cloudiness that is, the ITCZ location. Furthermore, the low-level convergence generated from pressure gradients associated with SST anomalies can feed into the regional and large-scale circulations (Lindzen & Nigam, 1987). Lamb (1978) demonstrated that large tropical Atlantic SST anomalies accompanied anomalous tropical Atlantic surface atmospheric circulation patterns that influenced the sub-Saharan rainfall. Specifically, Lamb (1978) found that warm SSTs anomalies ~300-500km south of 10°N and east of 35°W lead to droughts whilst northward anomalies lead to higher rainfall totals in the July – September period.

Fontaine and Janicot (1996) determined that drought events in West Africa are associated with warm SST anomalies in the eastern Pacific and the Indian Ocean, but drought specifically in the Sahel region corresponds to warm SST anomalies in the southern (equatorial) Atlantic and cold SST anomalies in the northern Atlantic. Flooding events in West Africa, to include the Sahel region, are associated with warming SSTs in the northern Atlantic. Fontaine *et al.* (2010) utilized observational, remotely sensed and reanalysis datasets to determine the impact of Mediterranean SSTs on West African rainfall and confirmed that warm events lead to rainier periods, but also warming of the Eastern Mediterranean affects the low-level wind flow such that it leads to a more northward migration of the West African Monsoon (WAM), an enhancement of the Tropical Easterly Jet and a decrease in the African Easterly Jet.

Giannini, Saravanan, and Chang (2003) identified the Sahel rainfall as being sensitive to SST anomalies and trends in the remote Pacific and Indian basins, and the local Atlantic basin. Giannini *et al.* (2003) further proposed that the warm equatorial Atlantic SST anomalies disrupt the WAM.

2.1.6 Land Use Changes

Deforestation in West Africa has been correlated with changing soil moisture, and thus latent heat fluxes, as well as sensible heat fluxes through the altered surface albedo over large areas. Researchers thought this to be linked to the rainfall variability experienced in the Sahel via the albedo-precipitation effect – the albedo of bare soil is lower than that of vegetation, thus the removal of vegetation over an area would increase atmospheric subsidence and reduce cloud formation creating a positive feedback loop as drier conditions would inhibit vegetation growth and increase rainfall perturbations south of the region (Charney, 1975; Otterman, 1974; Taylor, Lambin, Stephenne, Harding, & Essery, 2002; Xue & Shukla, 1993; Zeng, Neelin, Lau, & Tucker, 1999). Afforestation also influences the rainfall in West Africa via surface energy balance and cloud cover interactions (Xue & Shukla, 1996). Eltahir and Gong (1996) propose a relationship between vegetation and regional circulations including the West African Monsoon (WAM). Zeng *et al.* (1999) argued that the natural vegetation interaction influences the interannual variability and were able to mostly replicate the observation record, both in terms of magnitude and timing, with a coupled atmosphere-land-vegetation model that was driven by sea surface temperatures (SSTs). Taylor *et al.* (2002) showed that the change in land use in the Sahel prior to 1996 as a result of agricultural practices lead to a conversion of 4 percent of vegetated land to bare soil area. This change reduced rainfall totals annually, initiated the WAM a month too early, delayed the onset of the wet season in July (although the August totals were

unaffected), and did not allow the band of maximum rainfall to penetrate northward into the Sahel as observations indicate. In general, land use on its own as a driver of the dry spells observed in the 1990s in the Sahel is not sufficient, but cannot be discredited in climate modeling (especially regional climate modeling).

2.1.7 Synoptic-scaled Circulations – African Easterly Waves

Intraseasonal fluctuation of tropical convection occurs on short temporal scales, and is associated with synoptic-scale (large scale) wave disturbances. Over West Africa, these wave disturbances are referred to African Easterly Waves (AEWs). AEWs are wave-like disturbances of wavelengths of ~3000km, that propagate westward from northern Africa into the Atlantic Ocean every ~3-4 days during the boreal summer (Burpee, 1974). They are generated by the decreasing potential vorticity towards the north that is, baroclinic –barotropic instabilities within the African Easterly Jet (AEJ). AEWs can be characterized by low-level anomalies to the north of the AEJ and be associated with large MCSs such as squall lines as they propagate westward (Pytharoulis & Thorncroft, 1999).

2.1.8 Mesoscale Convective Systems and Mesoscale Convective Complexes

Superimposed on these intraseasonal changes are diurnal and inter-diurnal changes – mesoscale convective systems (MCSs) that cover various temporal and spatial scales. In general, MCSs include organized thunderstorms that persist for a few hours, and squall lines that persist longer (Figure 2.5). Mesoscale convective complexes (MCCs) are a subset of large and well-organized MCSs. Maddox (1980b) identified these features as weather systems that persist for at least six hours, with an average area of at least 100 000 km^2, at an average cloud top temperature

of 241K, with a colder convective core representing at least 50 percent of the overall area, with average temperature of 221 K.

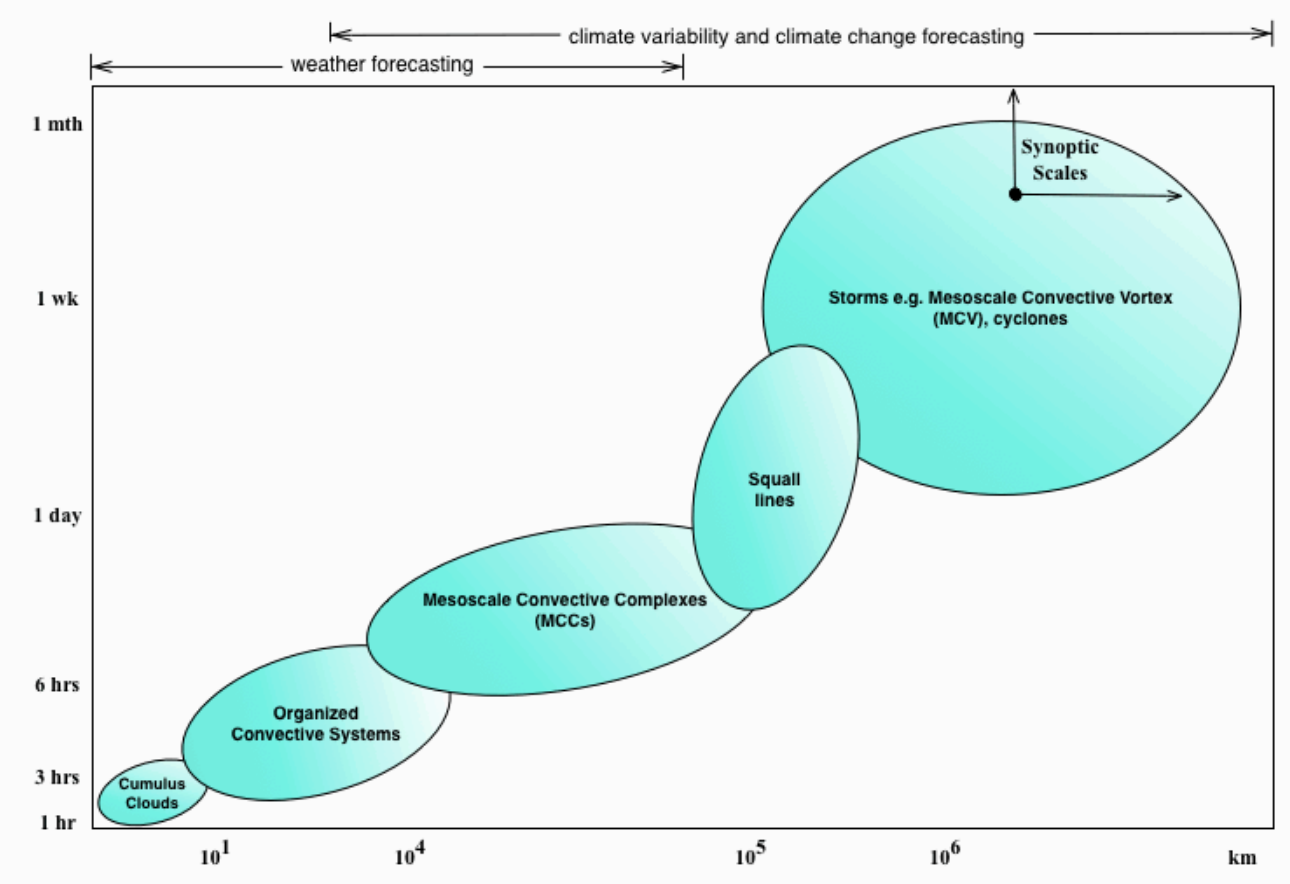


Figure 2.5. A schematic of the temporal and spatial scales of selected mesoscale convective systems.

The contribution of the 'short-lived' (24 hours or less) MCSs has been established. Le Barbé, Lebel and Tapsoba (2002), utilizing observation data from over 300 daily rain gauges covering an area of 1,700,000 km^2 in West Africa, determined that the number of the rainfall events is

highly correlated to the number of rainy days, where the ratio between the number of rain events and the number of rainy days depended on location and time of the year (Table 2.1). This indicates that an event occurring on a timescale less than daily is significantly contributing to the rainfall totals, and confirms an intraseasonal mode of the rainfall.

Table 2.1. The number of rainy days and the number of rain events observed in Niger between 1990 and 1998 (Le Barbé, Lebel & Tapsoba, 2002).

	May	Jun	Jul	Aug	Sep
Number of rainy days	5.9	7.8	11.4	13.9	7.7
Number of rain events	2.4	5.1	12.1	17.3	7.0

The high precipitation events during short periods associated with these events impact the climatological records of rainfall. For example, D'Amato and Lebel (1998) showed from a rain gauge network validated against infrared satellite imagery, that MCCs contribute to more than 70 percent of the seasonal rainfall in Niger, and MCSs more than 90 percent of the annual rainfall. They attempted to illustrate the relationship of these small-scale features in storm rainfall statistics either in the large-scale north-south propagation of the ITCZ or as random extreme events. They found that the mean storm rainfall in their study area was climatologically stationary, thus making it a stable parameter for characterizing rainfall in the Sahel, as it indicates the interannual rainfall variability is dependent on the number of rainfall events (MCCs). Laing, Fritsch and Negri (1999) considered the rainfall contributions from various satellite datasets during the summer Sahelian rainfall months and determined that MCCs

contributed ~22% of the rainfall. These findings indicate an understanding of the lifecycle, characteristics, frequency and seasonality of convective features globally, is important for several climate-related studies based on predictions of rainfall, such as hydrological studies, agriculture planning and disaster management.

The relationship between the regional circulations and the intraseasonal variability of MCS events (and thus rainfall totals) appears to be linked to the spatial scale of the MCS features. Mathon, Laurent and Lebel (2002) considered the relationship between the synoptic large-scale circulations (easterly waves) and small-scaled MCSs (defined as features lasting at least 3 hours with an embedded area greater than 5,000 km^2 at a temperature 213 K, and a mean speed greater than 10 ms^{-1}) over the Sahel from satellite images, reanalysis data and a network of 30 gauges for five rainy seasons between 1990–1994. They found that at the seasonal scale, the frequency of the sub-category of MCSs considered was not affected by the easterly wave occurrences, thus indicating no association to the interannual variability of rainfall. Velasco and Fritsch (1987) clearly established the relationship between MCC formation – a sub-category of large-scaled MCS – and the regional circulations such as the magnitude of wind shear, low-level jets, and latitude.

From the information provided, it is not far concluding that climate variability and climate change could influence MCC formation and frequency. However, known studies do not exist. Further, as seasonal rainfall totals are not only correlated with regional circulations, cloudiness and the number of large-scaled MCSs embedded in these regional features, but also with the rainfall (characteristics) associated with these large-scaled MCSs, studies related to them are merited to add to the understanding of convection and the processes that lead to convection

organization at the various mesoscales. Hence, this research will focus on mesoscale convective complexes – the large-scale sub-category of the mesoscale convective systems.

2.2 A Summary of Mesoscale Convective Complexes

2.2.1 Global Mesoscale Convective Complexes

Mesoscale convective complexes (MCCs), a sub-category of mesoscale convective systems (MCSs), are large, mesoscale, convectively-driven weather systems that are generally associated with high precipitation events during short periods (Maddox, 1980b). An example of the feature is provided in Figure 2.6 and the full definition is provided in Table 2.2.

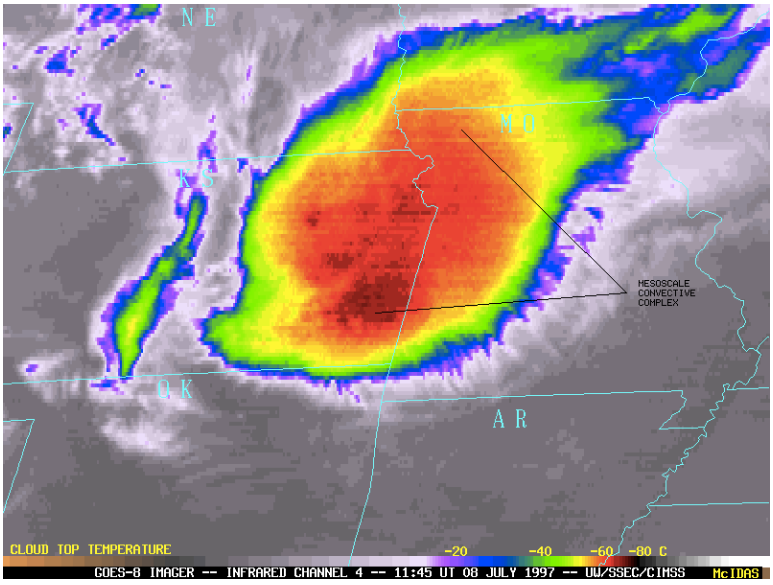


Figure 2.6: An example of a mesoscale convective complex (MCC) over the United States of America on 8 Jul 1997 from enhanced infrared satellite imagery.

Table 2.2. Mesoscale convective complex (MCC) properties based on infrared satellite imagery showing Maddox (1980b) criterion and the Laurent *et al.* (1998) variation of the criterion (in blue).

	PHYSICAL CHARACTERISTICS
SPATIAL RESOLUTION	A – Cloud shield with continuously low IR temperature ≤ 241 K must have an area ≥ 100 000km^2 B- Interior cold cloud region with temperature ≤ 221 K must have an area ≥ 50 000km^2 A – Cloud shield with continuously low IR temperature ≤ 233 K must have an area ≥ 80 000km^2 B- Interior cold cloud region with temperature ≤ 213 K must have an area ≥ 30 000km^2
MAXIMUM EXTENT	Contiguous cold cloud shield (IR temperature ≤ 241 K) reaches maximum size
SHAPE	Eccentricity ≥ 0.7 at time of maximum extent
DURATION	A and B must be met for a period ≥ 6hrs and ≤ 24hrs* *24hrs – there is no set maximum time in the literature however, it is recognized that if such a feature persists for more than 24hrs, it falls within the synoptic scale. At the same time, even though the temporal requirements may have been met for a synoptic system, the spatial requirements may not have been met.

MCCs are significant not only because of their ability to generate high amounts of precipitation in short durations on the weather timescale, but also the extent to which they alter the atmosphere and participate in the global energy and water budgets on the climate timescale (Duvel, 1989; Fritsch, Kane & Chelius, 1986; Gamache and Houze Jr., 1983; Leary & Houze Jr., 1980; Tollerud & Collander, 1993).

Laing and Fritsch (1997) studied the characteritics of MCCs globally. More than 400 MCCs occur at various locations around the globe, but mostly in tropical and mid-latitude locations, with approximately 66 percent of the occurrences in the Northern Hemisphere. They also are found mostly over continents though they can occur over oceans. The average duration of these systems is 10 hours, but this varies according to location. Generally, the systems are longer lasting in the Southern Hemisphere than in the Northern Hemisphere, and MCCs originating over the oceans are longer lasting originating over land. MCCs typically have cold cloud shields sized between 2–4 x 10^5 km^2, tending towards the larger sizes in locations over oceans and in the Southern Hemisphere. Studies of MCCs have been conducted over individual locations, in an effort to determine MCC characteristics and formation at that location. Examples include Velasco and Fritsch (1987) considering MCCs in the Americas and Blamey and Reason (2012) considering MCCs in Southern Africa. Though there are many similarities, there are also distinctions associated with location. Similarities regarding formation (in space and time), longevity and size, and are correlated with both the time of the year and day. MCC formation is favored in the summer season near the summer solstices in either hemisphere. Their frequency is dominated by the Earth-Sun relationship and the large-scale circulation pattern regarding the migration of semi-permanent high and low pressure systems. The summer changes of baroclinic atmospheric properties such as deformation and shear, especially in mid-latitude locations, create

a favorable environment for growth and sustaining their structure (Laing & Fritsch, 1997). Also, globally, MCCs tend to initiate in locations where there is: (1) a low-level jet; (2) a low-level convergence; (3) upper-level divergence; (4) an approaching mid-level vorticity maximum (associated with a weak short wave trough) (Laing & Fritsch, 2000).

Diurnally, MCC genesis occurs in the early afternoon and dissipation occurs in the early morning. Differences are related to the time of initiation, duration and cloud pattern characteristics. MCCs in the Southern Hemisphere initiate earlier and end later than those in the Northern Hemisphere by approximately two hours. Additionally, there are characteristic cloud patterns associated with the stages of development For example, MCCs in the Central Plains of the United States are often more linearly-shaped in their initial stages of cloudiness and cloud shields (Laing & Fritsch, 1997). MCC nocturnal organization is associated with differential radiative heating between the convective system that is warm core, and the environment that is cooler. In mid-latitude locations, this configuration is as a result of cloud-clear radiative processes whereby such temperature gradients lead to subsidence in the ambient atmosphere and low-level convergence into the system (Gray & Jacobson Jr., 1977). Additionally, radiative cooling at the cloud top and warming at the cloud base enhance the convective instability in mature convective systems (Webster & Stephens, 1980). In tropical locations, the convective systems are modulated via large-scale atmospheric destabilization (Miller & Frank, 1993). In general, large scale factors such as the position of long waves such as Rossby waves, and gradient zones between maxima and minima outgoing longwave radiation influence MCC formation (Augustine & Howard, 1991; Laing & Fritch, 1997).

MCC formation is also correlated with location on the globe. Regional factors influencing MCC formation are the position of the low-level jet, and low-level frictional effects associated

with elevation as MCCs tend to form on the lee-side relative to the prevailing mid-level flow of elevated terrain (Laing & Fritsch, 1993).

MCCs contribute to the local and regional mass, moisture and heat fluxes (Leary & Houze Jr., 1980, but their formation and variability is also correlated to large-scale circulations that are also influenced by mass, moisture and heat fluxes. As such, intuitively MCCs should have some inter-annual variability. However, there have been few studies focused on addressing this issue, in part due to the limited methods for evaluating long-term datasets of MCCs (methods for evaluating MCCs will be dealt with in the following section). Nonetheless, the few studies that have been conducted confirm there is some inter-annual variability of MCCs and MCSs in general. To illustrate, Velasco and Fritsch (1987) showed that MCCs in mid-latitude locations in South and Central America doubled in frequency for the 'El Niño' year compared to non-ENSO event for the period considered in their study, May 1981 – Apr 1983.

2.2.2 West African Mesoscale Convective Complexes

Laing and Fritsch (1993), Jobard and Desbois (1994), D'Amato and Lebel (1998), Laurent, D'Amato and Lebel *(*1998), Mathon and Laurent (2001), and Mathon, Laurent and Lebel *(*2002) have all explored the characteristics and rainfall contributions of MCCs and large-scale MCSs in the Sahel region, and West Africa in general Approximately 41 MCCs occur annually in West Africa in the boreal summer months July – September (Laing, Fritsch & Negri, 1999). Similar to the global population, West African MCCs are predominately formed and found over land – ~95 percent, with most activity in the Sahel region and little in the tropical rain forest belt of equatorial Africa (Laing & Fritsch, 1993).

The West African MCC development varies according to location. In central and North Africa regions between 5°S and 18°N – an area downstream of mountain ranges – generation is

favored due to the orography. These mountain ranges extend above 1.5 km and persistent vortices associated with them are observed in wind fields even at the 850 mb level during summertime (Laing & Fritsch, 1993; Viltard, Laurent & De Felice, 1990). As one approaches the coastline of the West Africa region (and the equator), MCC generation is associated with monsoonal effects of a surface thermal trough (Viltard *et al.*, 1990).

The West African MCCs, like the global population, are nocturnal features. Their diurnal cycle is similar to global populations, with thunderstorms developing in the mid-afternoon (~1400LST), MCC genesis occurring around 1900LST and reaching a maximum extent at ~0100LST then decaying at ~0600 – 1100LST. Their durations are longer than the global population, being ~11.5hours (Laing & Fritsch, 1997). In terms of the cloudiness, an elongation of the cold cloud shield characterizes the decay stage of these MCCs (as opposed to North American MCCs where the elongated shield characterizes the initation). West African MCCs' typical sizes are smaller than the global average between 2–4 x 10^5 km^2, and their cloud shields colder ranging between 240K and 213K. Laurent, D'Amato and Lebel (1998) determined smaller areas according to the colder temperature ranges associated with the cloud shield. The cloud shields associated with convective cloud in this region are within the temperature range 233K – 213K, and the area between 3–8 x 10^4 km^2. The temperature ranges for West African MCCs are generally colder than the Maddox (1980) criterion as they are adjusted to accommodate the deeper tropical atmosphere (Jobard & Desbois, 1994). West African MCCs propagate westward at ~12–15 ms^{-1} towards areas of high equivalent potential temperature over the warm oceans (D'Amato & Lebel, 1998; Lebel, Taupin, & D'Amato, 1997).

2.2.3 Identifying Mesoscale Convective Complexes

Mesoscale convective complex (MCC) identification (and identification of large-scale mesoscale convective systems (MCS) features) utilizes a spatial-temporal criterion that requires high-resolution (30-minute to 3-hourly) geostationary weather satellite data (Maddox, 1980a,b). Specifically, MCC identification utilizes infrared (IR) imagery and visible (VIS) imagery. The IR satellite data provides the variable brightness temperature (T_B), which provides information about the temperature of an object from which radiation is emitting, and also infers information about the altitude of the object. In general, the higher an object is from the surface, the colder its T_B. Further, cold T_B indicate cold cloud tops, with colder temperatures indicating clouds with vertical extent and convection. Desbois, Kayiranga, Gnamien, Guessous and Picon (1988) demonstrated that the 10.8 μm IR channel is good for monitoring deep convection, as the data from this channel represents the effective cloud top temperature (T_B). The VIS data is analogous to a black and white photograph of the Earth's surface from space, and provides a means to validate the cloud types identified from the temperature data through analyzing shape characteristics of clouds. This method is preferable as the spatial expanse of MCCs and variability in their convective characteristics make them difficult to characterize even in dense networks of radars and / or surface gauges.

2.2.3.1 The Mesoscale Convective Complex criterion

Maddox (1980b) provides a criterion for identifying mesoscale convective complexes (MCCs) in mid-latitudes based on spatial resolution of the cloud shield at a given temperature, the shape of the system, and the duration of the system (Table 2.2). An important aspect of the MCC identification criterion is the shape, defined as the eccentricity at the time of maximum

extent of the feature. The MCC is a quasi-oval / quasi-circular shape, the eccentricity is expected to be greater than 0.7. This is important because features that meet the other aspect of the Maddox (1980b) criteria, but not the shape criteria, are categorized as another MCS feature, such as squall lines (Anderson & Arritt, 1998)

Several modifications of Maddox (1980b) criterion for MCCs in West Africa were made. These include a version to account for a higher temporally resolved dataset by Laing and Fritsch (1993). Additionally, Houze Jr. (1993) produced a modified version that contains an additional characteristic stating that the cloudy area defined produces a contiguous precipitation area of a minimum 100 km in at least one direction Laurent, D'Amato and Lebel (1998) produced a version that reduces the spatial resolution thresholds to A – cloud shield with IR temperature ≤ 233 K and an area ≥ 80,000km^2, and B – cloud shield with IR temperature ≤ 213 K and an area ≥ 30,000km^2, to allow for better depiction of convective rain-producing clouds during the rainy season in the Sahel. Recent studies such as Goyens, Lauwaet, Schröder, Demuzere and Van Lipzig (2011), utilized the Laurent *et al.* (1998) criteria, making it feasible for this study (Table 2.1).

2.2.3.2 Methods to Identify Mesoscale Convective Complexes

Studies of MCCs utilized manual, automated and / or semi-automated methods to identify MCCs according to a criterion. Approaches to these methods can be defined as Eulerian or Lagrangian (Schröder, König, & Schmetz, 2009). In the Eulerian approach, the areas of deep convection on the images are retrieved and stored as a function of spatial position according to the local solar time. A harmonic analysis is then applied on each spatial position within the time bins. In the more commonly used Lagrangian approach, the convection is tracked by considering an area of a given temperature overlap between successive images.

Initially the methods for identification were fully manual, implemented a Lagrangian approach and required both infrared (IR) images that supplied the cloud-top brightness temperature (T_B) for providing cloud characteristics, and visible (VIS) images that verified that information. This was illustrated by both Velasco and Fritsch (1987) and García –Herrera *et al.* (2005). These manual method studies identified the characteristics of the MCC, their cold cloud shield areas and temperature, their duration and their lifecycle. However, the method is subjective, time-consuming and inconsistencies between observers in identification are likely.

The manual method evolved into semi-automatic methods that incorporated a human factor to corroborate the automatic implementation using VIS images in studies as seen in studies from Laing and Fritsch (1993), Blamey and Reason (2012). Automated implementations that do not require VIS imagery also evolved as seen with Carvalho and Jones (2001), Schröder, König, & Schmetz (2009), Vila, Machado, Laurent, and Velasco (2008). Automated methods employ algorithms to search time-series of infrared satellite images to identify systems. Although time-efficient (in terms of human resources), this method can falter with the identification of splitting and merging systems between consecutive images. Semi-automated methods use computer programs to identify systems, and then manually corroborate, utilizing both IR and VIS images. The studies that implemented semi-automatic and automatic methods added to the knowledge base regarding the characteristics of MCCs according to location. Additionally, the studies started to explore the precipitation characteristics of MCCs through considering other IR bands and microwave bands that the satellite sensors provided. However, most of the existing and generally utilized semi-automatic and automated methods, such as Carvalho and Jones (2001) and Schröder *et al.*, (2009), are based on some forward and backward in time algorithm that tends to be inefficient in terms of computer resources. More importantly, these methods often

falter when evolving cloud systems that may merge or split over time, and, thus, still require human invention.

More recently, a fully automated method, called the Tracking of Organized Convection (TOOCAN) algorithm, has been introduced for identifying and tracking MCSs in meteorological applications (Fiolleau & Roca, 2013). This method is based on clustering of similar brightness temperatures in sequential geostationary satellite images. The objective of the TOOCAN algorithm is to identify MCSs by associating areas of coldest brightness temperatures (convective seeds / cores) in sequential IR data through a clustering method that identifies the convective seeds (areas of deepest convection) according to a T_B maximum threshold, then iteratively over temperature ranges (for example, 1 K variation) track the MCS from this radial point outwards in a 3-dimensional space (time, latitude and longitude) as a MCS. The TOOCAN algorithm addresses the concerns of previous MCSs tracking methods and adds value to meteorological applications involving MCSs, as it provides information about the core structure of the system, but although no algorithm analysis was provided intuitively the algorithm does not appear efficiently scalable to long-term datasets.

Automated methods generally have two distinct stages: (1) the cloud detection or cloud masking stage, and (2) the tracking or evolution stage. The cloud detection or cloud-masking stage involves reducing a full IR satellite dataset to only the cloud areas of interest, by imposing temperature, and area or volume restrictions. The tracking stage evolves small MCSs into MCCs.

The cloud detection stage is defined by imposing a temperature-area (or volume) criterion that varies according to the particular MCS and the location study. (Note that the TOOCAN algorithm does not impose a temperature criterion.) The temperature ranges from a 255 K

temperature threshold to identify precipitating areas in large MCSs, to a 195 K temperature threshold to identify the convectively most active parts of the system (Desbois, Kayiranga, Gnamien, Guessous & Picon (1988); Duvel, 1989; Fiolleau & Roca 2013; Machado, Rossow, Guedes, & Walker (1998); Mapes & Houze, 1993). The area ranges from 2400 km^2 to 5000 km^2 for MCC development (and large MCSs) (Machado & Laurent, 2004; Mathon & Laurent, 2001; Vila *et al.*, 2008). Further, smaller area cut-offs tend to produce too many cloud areas from the cloud masking area process that are of no consequence to the MCC identification. Mathon and Laurent (2001) showed that cloud areas (MCSs) of size larger than 5000 km^2 and temperature thresholds of 253 K, 235 K and 213K did not affect the total number of tracked MCCs in West Africa. For the same location, Machado and Laurent (2004) successfully used a cut off area of convection, sized 3500 km^2 with a T_B of at least 233 K, for successive images a maximum of two hours apart, while Goyens, Lauwaet, Schröder, Demuzere & Van Lipzig (2011) used the same size for images 30 minutes apart. As such, the minimum cloud area criterion required to evolve into a MCC in West Africa is an area 3500 km^2, with a brightness temperature of at least 233 K. Problems have been identified with these approaches from a meteorological (short time scale) perspective. The issues include temperature thresholds may not delineate the edges of the cloud systems clearly, area cutoffs may discredit valid systems, the convective portion of the cloud may not necessarily be identified, and irrelevant data storage (Chaudhuri & Middey, 2009; Chaudhuri & Middey, 2011).

The tracking approaches existing within the literature, include maximum spatial correlations as illustrated by Carvalho and Jones (2001), a Lagrangian approach based on projected centroid location as outlined by Johnson *et al.*, 1998, propagation speed of criterion as shown by Woodley, Griffith, Griffin, and Stromatt (1980) and Machado, Rossow, Guedes, and Walker

(1998), maximum overlap area as demonstrated by Laurent, D'Amato, and Lebel (1998), a global cost function to enforce shape and path characteristics, pointed out by Chaudhuri and Middey, (2011), and greedy optimization of overlap areas as exemplified by Dixon and Wiener (1993). For example, in West Africa, Machado *et al.* (1998) determined the speed between MCCs of successive images be no more than 19.5 ms^{-1}. The maximum spatial correlation involves finding the region of maximum spatial correlations between successive images, whereas the area-overlap method considers the maximum overlap of areas between the successive images. As an MCS evolves overtime, a number of evolution stages can occur between successive IR images (Figure 2.7). Specifically, systems may increase in size (grow), decrease in size (decay), remain the same size (maintain), become more than one system (split), amalgamate two or more MCSs (merge), or a new feature may appear. Machado *et al.* (1998) determined the overlapping tracking method efficiently evolves systems based on these characteristics but with careful consideration for the time between successive images (three hourly images) and the parameters to track the systems. This area overlapping tracking approach is thus more efficient than the spatial correlation method, which usually falters with merging and splitting systems (Blamey & Reason, 2012). However, the most common method used in automated tracking algorithms is the area-overlapping method (Schröder, König, & Schmetz, 2009; Goyens, Lauwaet, Schröder, Demuzere, & Van Lipig, 2011).

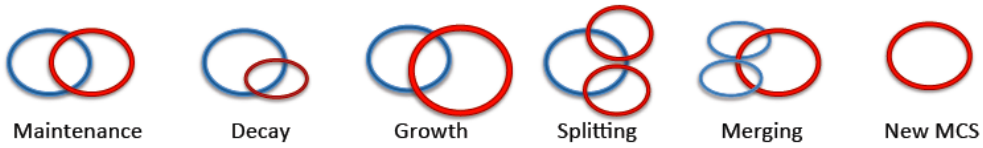


Figure 2.7. Schematic of the possible fates of MCSs between successive infrared satellite images.

In general, the existing semi-automated and automated methods are based on forward-backward in time algorithms as sequential data images are considered in the tracking stage. Maximum Spatial Correlation Tracking Technique (MASCOTTE) algorithm, which is implemented in Interactive Data Language (IDL), evaluates the evolution of one MCS (for example, one MCC) throughout the time of the dataset at a given iteration (Carvalho & Jones, 2001). This approach requires large computer memory to retain the entire data and copies of the dataset, whilst the program is operating. Thus, for large areas or long-term datasets, this can become a very computationally expensive method. The Schröder, König, & Schmetz (2009) method is based on an overlapping of 5 percent between successive images, better allowing for splitting and merging of events. The algorithm commences by identifying areas of interest via the cloud masking stage. Once these areas have been identified, the algorithm backward-steps to determine if a given contiguous area meets the overlap criteria. If it does, it continues to step backwards in time until the minimum criteria for deep convection is not found. Once this is found, the algorithm then checks forward in time for the development of the system until a 5 percent overlap cannot be found. It then stores the MCC information and moves on to the next one in the domain (a brute force approach). In general, from the existing methods, it is demonstrated that a method that efficiently handles the data for long-term records is required.

2.2.3.3 Determining the Rainfall Amounts Associated with Mesoscale Convective Systems

Quantifying the amount of surface precipitation associated with a particular mesoscale convective system (MCS) or mesoscale convective complex (MCC) is challenging, given the small spatial and temporal resolution of the feature and thus the dense rain gauge network that would be required. Precipitation derived from satellite data, such as infrared and microwave data, can help determine cloud characteristics and precipitation rates of features in West Africa. Jobard and Desbois (1992) demonstrated that precipitation estimates from convective features based on IR images alone are insufficient as infrared radiances alone provide information about the cloud cold top temperatures but not the precipitation. They further demonstrated that rainfall estimates from the combination of IR and microwave data was more accurate for the West Africa location. Nicholson *et al.* (2003a) found that IR and microwave datasets displaced the rainfall belt northward of the gauge measurements that were taken as truth for the period June-July-August 1988–1994. Furthermore, the rainfall captured by the microwave-based data was unable to capture the rainfall maximum associated with climatological features such as orographic rainfall. Nonetheless, Laing, Fritsch and Negri (1999) determined the average rain area associated with Sahelian MCCs is 285,000 km^2, with an average rainfall of 34 mm, through a relationship they established between microwave-derived precipitation data and IR data.

A potential satellite dataset that can overcome the limitations of rainfall estimates as outlined is the NASA Tropical Rainfall Measuring Mission (TRMM) dataset, which provides the distribution of rainfall within the tropics through combining data from various satellite instruments[1]. The TRMM composite dataset has been available since 1997. Nicholson *et al.*

[1] http://trmm.gsfc.nasa.gov/

(2003b), using TRMM data, found that the TRMM (3B43) merged datasets formed from combining the data from the instruments, performed well showing excellent agreement with gauge data over West Africa on monthly to seasonal timescales and at the 25km spatial resolution, but individual components such as the precipitation radar (PR) or the TRMM microwave imager (TMI) over-estimated rainfall and did not perform well.

TRMM data can also be used to monitor the vertical structure of precipitation features as illustrated by Jenkins (2000). Nesbitt, Cifelli, and Rutledge (2006) used satellite data, specifically TRMM version 6 data, to associate the vertical and horizontal structure of precipitation features (including MCCs) in the tropics with rainfall characteristics such as precipitation rates. They used visible and infrared scanner (VIRS) and TMI reflectivity to identify brightness temperatures and correlated with PR and Lightning Imaging Sensor (LIS) data. They found the VIRS and TMI data provided larger, more uniform coverage whereas the PR and the LIS composite data provided better demarcations of the heavy precipitation. Nukilin *et al.* (2012) determined that for hourly resolved rainfall, there was no value added by using microwave data verses IR data within the TRMM data.

Isolated studies allowing for distinctive features regarding MCCs have also been resolved via satellite data. For example, Goyens, Lauwaet, Schröder, Demuzere, and Van Lipzig (2011) determined from METEOSAT-8 IR images and TRMM, that large-scale MCSs features in the Sahel region are approximately 57,000 km^2 (duration nine hours with an embedded core lasting six hours) with a precipitation peak of ~12.3 mm hr^{-1}. Goyens *et al.* (2011) utilized an automatic Lagrangian MCC method in their study, specifically the method employed by Schröder, König and Schmetz (2009), and manually compared the identified MCCs with similar areas in TRMM

data. Currently, there are no automated methods for determining precipitation characteristic of MCSs or MCCs from satellite data.

2.3 Data mining, management and distribution

Data mining derived out of the "data rich, but information poor" era and refers to the process of discovering patterns in big data to extract data and transform it to an understandable structure, for example a relational database or a graph, for taking out information (Han, Kamber & Pei, 2006). Information retrieval has lead to data distribution and management related problems, for example, issues related to how to store data in databases for ease of access and retrieval. One implementation to this data management and distribution problem is the Apache Object Oriented Data Technology (OODT) which is a framework of distributed objects and databases that allows for seamless management, manipulation, analysis, and visualization of distributed stored data (Mattmann *et al.,* 2009). For example, in the climate science application of the Regional Climate Model Evaluation Database (RCMED), Apache OODT employs other ASF projects, namely Apache Hadoop (White, 2012), Apache HIVE (Capriolo, Wampler & Rutherglen, 2012), and Apache Sqoop (http://sqoop.apache.org/). The required services from Apache Sqoop for efficiently and effectively transferring of bulk data between structured data stores, and Apache Hive for facilitating efficient processing data that is distributed between data stores, are wrapped up in Apache OODT's framework (Mattmann *et al.*, 2013).

Data mining activities have evolved into graph mining where graphs are used as the structure of data mining activities. A graph is similar to a network with a number of points, called vertices, and a number of lines, called edges (Trudeau, 1993). The complexity of a graph is determined by the number of vertices and edges between them, within the graph. Graph mining – data mining leading to graph formation – and distribution is increasing in popularity in various fields leading

to numerous graph designs and complexities of implementation (Aggarwal & Wang, 2010). For example in chemical application, atoms may be considered as nodes and bonds as edges. As the application varies, the underlying mining algorithm changes, hence from a strictly computer science algorithms design perspective, this has become a quickly growing field of interest.

CHAPTER 3. DATASETS AND TOOLS

Satellite (remotely sensed) datasets are commonly used in studies contributing to current scientific understanding and prediction of rainfall characteristics and variability, as these datasets provide a reasonable spatial and temporal resolution of the observed precipitation characteristics. Such datasets can be seen in works by Adler, Negri, Keehn, and Hakkarinen (1993), Smith, Mugnai, Cooper, Tripoli and Xiang (1992), Mugnai, Smith and Tripoli (1993), Mathon, Laurent and Lebel (2002) and Goyens, Lauwaet, Schröder, Demuzere, and Van Lipzig (2011). Satellite datasets are inherently large as they cover large spatial areas and are generated frequently. As such, satellite datasets fall into the category of big data – datasets that are voluminous, contain various formats and are required at high speeds (Jacobs, 2009). Arnaud, Desbois and Maizi (1992) highlighted the big data issues related to using long term records of satellite data for identifying mesoscale convective systems (MCSs). Two decades later, the big data issue continues to be a hindrance and has escalated with the increased number of satellite missions and model data, which is sometimes used in retrieval methods and / or composite datasets, thus contributing to the sum of data (Mattmann, 2013; Overpeck, Meehl, Bony, & Easterling, 2001).

3.1 Satellite Datasets for Characterizing Mesoscale Convective Systems in West Africa

The data from each type of weather satellite dataset provides a set of properties that allow for distinguishing between cloud forms and / or morphology. As aforementioned, the identification of mesoscale convective complexes (MCCs) (and mesoscale convective systems (MCSs)) depends on infrared (IR), and sometimes visible (VIS), satellite data. Details about the rainfall associated with MCCs (and MCSs) depend on IR and Microwave (MW) satellite dataset.

Satellite datasets are referred to as remote sensed data. Remote sensed data involves the interpretation and inversion (calculation) of radiometric measurements of electromagnetic radiation (Figure 3.1) that were measured some distance away (in this case by a sensor onboard a satellite). Remote sensed data can be further categorized into active and passive remote sensing. Active remote sensing uses an artificially generated source of radiation to determine the attenuation of the beam between the transmitter (where the radiation originates), and the detector (the sensor that reads the radiation). Passive remote sensing uses natural sources of radiation from the Sun or radiation from within the climate system. In both active and passive sensing, attenuation of the electromagnetic radiation beam may occur due to one or a combination of the transmission, scattering, reflection and absorption.

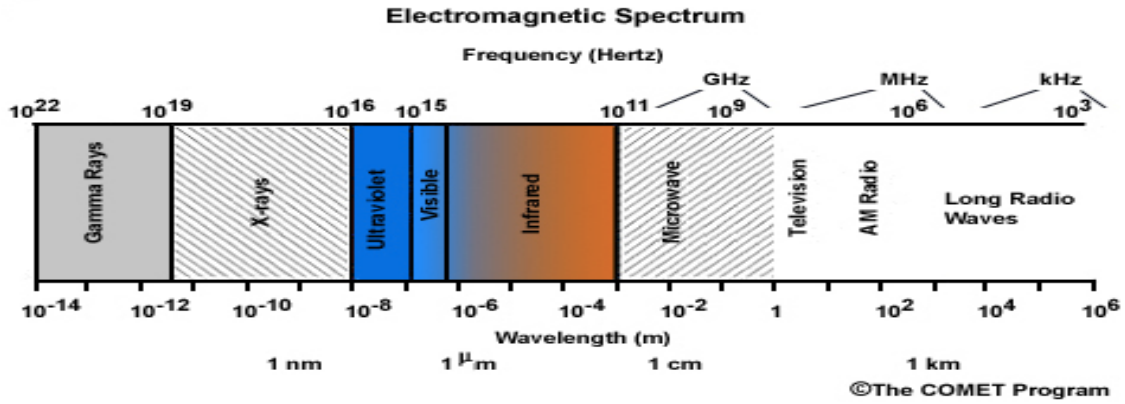


Figure 3.1. The Electromagnetic Spectrum (Laing and Evans, 2011).

The wavelengths at which the sensors onboard the satellite measure the radiation are analogous with the dataset. In atmospheric studies, the range of radiation, from the ultraviolet through to the microwave wavelengths, is useful. Furthermore, infrared and microwave sensed

radiation are most useful in this study. The reader is directed to Liou (2002) Chapter 7 and Laing and Evans (2011) for further details on satellite retrieval methods.

3.1.1 Satellite Data for Identifying Cloud Characteristics – Infrared Satellite Data

The infrared (IR) satellite data provides information about the extinction of the emitted IR radiation by a medium (the surface), assuming a non-scattering environment in local thermal equilibrium. IR sensors usually fly onboard geostationary satellites at a height ~36,000 km above the Earth's surface, so as to maintain a fixed view of an area (~42 percent of the globe) at a time, allowing for continuous monitoring of that location. IR sensors also employ passive remote sensing methods. The satellite sensor detects within the thermal band (10-12μm) of the electromagnetic spectrum, and observes downwards to the surface (or as far as possible) in a narrow cone, with the tip of the cone originating at the local vertical of the instrument. The IR sensor observes a particular location continuous in space and time, irrespective of the time of the day. IR data provides the variable brightness temperature (T_B) that gives information about the cloud temperature and altitude, where, as a general rule, the colder the object, the larger the altitude. T_B is uniquely related to the (sensed) wavelength.

The weather IR satellite instruments provide information about the cloud top temperature, irrespective of the time of the day, as IR radiation is consistently being emitted by the Earth-atmosphere system. As the intensity of the IR radiation reaching the satellite sensors depends on the warmth of the object emitting that radiation, interpretation is such that most intense radiation comes from warmer objects such as the Earth's surface, while cooler objects such as clouds emit less radiation. The satellite image may be represented as grayscale (warm areas are darker, colder areas are whiter) or false color. In general, interpretation of IR satellite data provides further information about the cloud type, shape, morphology and other characteristics. Additionally, the

limitations of visible imagery (VIS) are addressed in IR images. Hence, many studies have used the IR data in identifying cloud systems. However, within the IR dataset, the limitation of the wavelength dependence of T_B for studies exists.

In weather applications, IR satellite data can be used for identifying and characterizing weather features, and estimating precipitation specifically from convective clouds, as convective cloud tops are higher, and thus correlate with high precipitation rates. Using IR data for precipitation estimation requires that the T_B are averaged over location and time, then compared to precipitation measurements (for example, rain gauge measurements). Though there are advantages to this technique, major limitations arise regarding the resolution of the satellite sensor being too coarse to capture convective-scale structures (as they develop or that embedded in mature systems), and inherently would not handle precipitation from warm clouds.

3.1.1.1 The Infrared Brightness Temperature Dataset Used

Desbois, Kayiranga, Gnamien, Guessous, and Picon (1988) demonstrated that the IR 10.8 μm channel is good for monitoring deep convection, as the data from this channel represent the effective cloud top temperature (T_B). Previous studies, such as Laing and Fritsch (1993) have utilized the METEOSAT IR images from the International Satellite Cloud Climatology Project (ISCCP) B3 stage radiance data. The METEOSAT is a geostationary meteorological satellite, operated by the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT), with multiple series. The instrument onboard for sensing is the Spinning Enhanced Visible and Infrared Imager (SEVIRI) that has the capacity to observe in 12 spectral channels – four visible and near IR channels, and eight infrared channels[1]. The IR images used for cloud detection are the IR 9.7, 10.8, and 12.0 μm. These images are temporally resolved

[1] http://www.eumetsat.int/Home/Main/Satellites/MeteosatSecondGeneration/Instruments/index.htm

every 30 minutes. Spatial resolution varied according to the series of METEOSAT, but is currently ~5 km.

The National Centers for Environmental Protection (NCEP), the Climate Prediction Center (CPC), and the National Weather Service (NWS) created a global dataset (between 60 °N – 60 °S) composite infrared dataset from data collected by sensors onboard various geostationary satellite platforms, including NASA GOES-8/10, EUMETSAT METEOSAT-7/5 and Japan's Geostationary Meteorological Satellite (GMS). The GOES-8/10 flies a five-channel imager onboard that dedicates one channel to visible wavelengths and the other four to infrared wavelengths. The channel used for detection of mesoscale features in synoptic disturbances is Channel 4 (10.2–11.2 μm spectral band). The GMS carries the Visible and Infrared Spin Scan Radiometer (VISSR), collecting data every 30 minutes in the visible 0.50–0.75 μm spectral band at 1.25 km resolution, and infrared 10.5–12.5 μm band at 5.0 km resolution. All the data from these various platforms are corrected for the varying zenith angle of each of the contributing satellite sensors and provided at a 0.036° or 4 km spatial resolution and 30-minute temporal resolution. Also referred to as MERG[2] dataset, it is the preferred dataset for this study.

The MERG dataset is available from February 7th, 2000 to present. The area of West Africa as defined by satellite images requires consideration of the curvature of the Earth's surface and viewing angles. Low viewing angles usually require large adjustments that can be potentially problematic within studies. Issues related to low viewing angles, such as the issue that Laing and Fritsch (1993) experienced with METEOSAT data for latitudes exceeding 20°, are avoided here through utilizing the MERG dataset (as the correction has been already made). For this study, hourly MERG data, on the hour, (that is ':00') will be utilized, as this temporal resolution is

[2] http://mirador.gsfc.nasa.gov/collections/MERG__001.shtml

sufficient for identifying and tracking large MCSs features (Goyens, Lauwaet, Schröder, Demuzere, and Van Lipzig, 2011).

3.1.2 Satellite Data for Precipitation Characteristics

3.1.2.1 Microwave Satellite Data

Microwave (MW) satellite sensors directly detect liquid and / or ice particles and precipitation within the range of size from 10 to 100 micrometers (μm) throughout the depth of the atmosphere, which immediately makes MW datasets superior to IR satellite inferred precipitation datasets. MW instruments fly onboard polar-orbiting satellites or low-orbiting satellites and employ both passive and active forms of sensing. Thus, the sensors cover the globe in swaths of a given width, repeating a swath area every ~12 hours. An example of the final product is provided for two MW sensors in Figure 3.2.

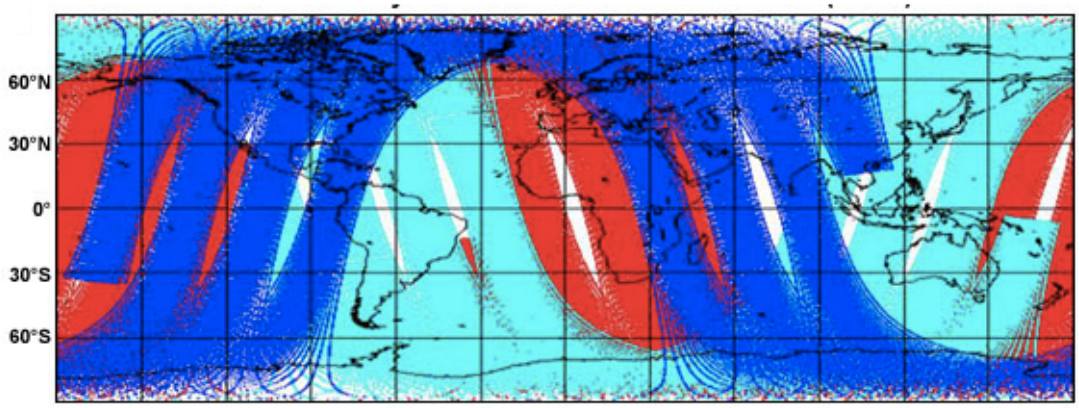


Figure 3.2. Global microwave coverage from sensors onboard various satellites whose swaths are represented by the different colors. The white areas indicate regions where no measurements are made (Laing and Evans, 2011).

The MW sensor cannot detect continuous throughout the range, hence several discrete MW wavelengths, which are usually described by their frequency in gigahertz (GHz) as opposed to wavelength, are chosen. These are referred to as channels (Figure 3.3).

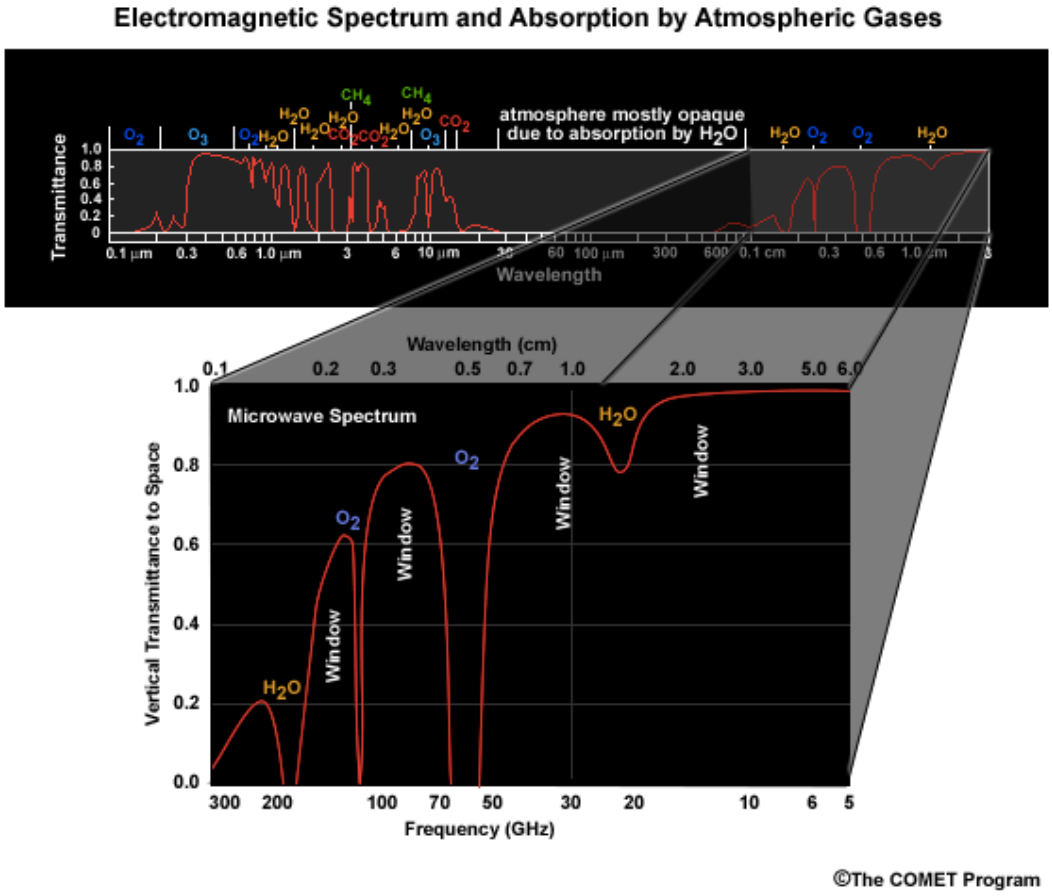


Figure 3.3. The electromagnetic spectrum highlighting the microwave portion and its absorption in the Earth's atmosphere (Laing and Evans, 2011).

On weather satellite instruments, channels are chosen such that some are located in the atmospheric window regions – areas where atmospheric gases absorb very little radiation, and some are located where there is high absorption – in the MW region of the electromagnetic spectrum only water vapor and molecular oxygen exhibit significant absorption (Figure 3.3). By doing this, the sensor can observe the surface without interference (even in the presence of physical objects such as clouds) in the window channels, and compare these results with the channels from high absorption to derive surface, cloud and precipitation properties.

Similar to infrared (IR) data, the characteristics that can be determined about cloud and precipitation in MW data vary according to the channel (radiation wavelength). In the 85–91 GHz channel, deep convection appears as a cold area, and water clouds and air masses with high water vapor content (but not yet clouds) are observed even over water surfaces, as areas with warm brightness temperatures. The data from the 85–91 GHz channel also provides information below thin cirrus clouds, and can distinguish deep convection areas, even though it may not be able to penetrate deep within the convective region to determine low-level structure as the MW is attenuated. The 37 GHz channel is a higher resolution channel to the aforementioned channel that shows water clouds and precipitating clouds as warmer against a colder ocean, and brings low-level cloud and rain features to the foreground as the measured upwelling MW radiation is mostly unaffected by ice particles. The 10 GHz channel provides detailed information about precipitation intensity, especially in locations of high precipitation rates.

The MW datasets provide a high spatial resolution but low temporal resolution of data, because the low-orbiting satellite platforms. To circumvent this shortcoming, some satellite precipitation estimates have combined MW measurements with estimates from IR, which provide a higher frequency.

3.1.2.2 The Precipitation Satellite Dataset Used in this Study

Satellite data from the Tropical Rainfall Measuring Mission (TRMM), - a mission between NASA and the Japan Aerospace Exploration Agency (JAXA), provide the distribution of rainfall within the tropics through combining data from various instruments[3], and is used in this study. TRMM was launched in 1997 on an equatorial orbit, staying between 35 degrees north and south of the equator. There are five instruments on board:

1. The TRMM precipitation radar (TRMM-PR): an active microwave sensor designed to support three-dimensional views of cloud and storm structure. The PR swath is 220 km;
2. The TRMM microwave imager (TRMM-TMI): a passive microwave remote sensor, with a 750 km swath that quantifies water vapor, cloud water and rainfall intensity;
3. The TRMM visible and infrared scanner (TRMM-VIR): a passive five-channel cross-track scanning radiometer in the visible and infrared spectrum;
4. A Cloud and Earth Radiant Energy Sensor (CERES): CERES instruments are passive remote sensors that measure both within the visible and infrared wavelengths. The original CERES that flew onboard the TRMM satellite only operated from January to August 1998, and March 2000;
5. The Lightning Imaging Sensor (LIS): a high-resolution passive remote sensor measuring lightning strikes in the visible spectrum.

The three-hourly TRMM 3B42 dataset is a merged high quality / IR precipitation and root-mean-squared precipitation error estimate. The dataset is produced from the multiple TRMM sensors and IR-precipitation estimates from sensors on other satellites. This composite dataset is prepared in two major and distinct steps. The first step combines the TRMM VIRS and TRMM

[3] http://trmm.gsfc.nasa.gov/

TMI (1B01 and 2A12 product, respectively) orbit data with the monthly TMI / TRMM Combined Instrument (TCI) calibration parameters (3B31 product), to obtain monthly IR calibration parameters. In the second step, the outputs from the first step are used to adjust the merged-IR precipitation data. The TRMM 3B42 version 7 dataset is available at a 0.25° (~25km) resolution every three hours. Combining the IR and the MW data in this product allows for better coverage of features. Table 3.1 provides a summary of the data characteristics of the datasets to be used in this study.

Table 3.1. A summary of the data characteristics of the remote-sensed datasets used in this study[4,5,6] . * indicates the actual variable to be analyzed.

TRMM 3B42v7 Data characteristics		
Characteristics	Temporal Coverage	1998-01-01 to Present
	Temporal resolution	3-hourly
	Geographic Coverage	50°S - 50°N, 180°W - 180°E
	Spatial resolution	0.25°x0.25° (nlat = 400, nlon = 1400)
Contents	Precipitation*	TMPA precipitation estimate. Units: $mmhr^{-1}$
	relativeError	TMPA random error estimate. Units: $mmhr^{-1}$
	HQprecipitation	Pre-gauge adjusted MW precipitation estimate at each 0.25° x 0.25° box. Units: $mmhr^{-1}$
	IRprecipitation	Pre-gauge-adjusted IR precipitation estimate at each 0.25°x0.25° grid box. Units: $mmhr^{-1}$
	satObservationTime	Satellite observation time minus the time of the granule of each 0.25°x0.25° box. Units: minute
Globally merged Full Resolution IR Brightness Temperature data (MERGv1)		
Characteristics	Temporal Coverage	2000-02-07 - present
	Temporal resolution	1-hrly (30 min)
	Geographic Coverage	60°S - 60°N, 180°W - 180°E
	Spatial resolution	0.036° x 0.036° (4km) (pixel resolution)
Contents	Brightness temperature*	Kelvin

[4] http://disc.sci.gsfc.nasa.gov/precipitation/Globally_merged_IR.shtml
[5] http://disc.sci.gsfc.nasa.gov/daac-bin/DataHoldingsPDISC.pl?LOOKUPID_List=MERGED_IR
[6] http://disc.sci.gsfc.nasa.gov/precipitation/documentation/TRMM_README/TRMM_3B42_readme.shtml/

3.2 The Regional Climate Model Evaluation System and the Apache Open Climate Workbench

Satellite datasets are somewhat difficult to use because they have varying spatial and temporal resolutions, and are stored in non-homogenous formats. Furthermore, satellite datasets are inherently large as they cover large spatial areas and are generated frequently, usually in the order of minutes. Thus, this project requires big data management methods.

Within the atmospheric sciences, big data management problems also occur with climate model evaluations. Predictions from climate models (global and regional) and observation data are necessary for informing decision-making processes related to climate impacts from both natural and anthropogenic climate change. However, before a climate model prediction can be used in the decision-making process, the model needs to be validated. A significant portion of validating models is to evaluate their outputs to a historical period of observed data to determine model biases

The Regional Climate Model Evaluation System (RCMES, http://rcmes.jpl.nasa.gov/) is a tool that was developed by National Aeronautics and Space Administration (NASA) Jet Propulsion Laboratory, California Institute of Technology (JPL), and their Joint Institute for Regional Earth System Science and Engineering (JIFRESSE) with the University of California, Los Angeles (UCLA) to inherently handle big datasets in the Earth Sciences. RCMES seeks to improve access to existing quality-assured long-term climate observations for the evaluation of regional climate projections through the inclusion of NASA remotely-sensed observations. RCMES allows for the evaluation of regional (and global) climate projections, the facilitation of seamless evaluations with datasets from various sources, and the promotion of better assessments of climate impacts, vulnerabilities and risks to natural and anthropogenic climate change for

various sectors (Whitehall *et al*., 2012). As the RCMES project evolved, aspects were donated to the Apache Foundation through the Open Climate Workbench project (Apache OCW, http://climate.apache.org/). The Apache OCW and RCMES project have evolved such that Apache OCW has developed and made available the inner methods of RCMES in a succinct modular format that identifies the purpose of various functions as they relate to big four dimensional datasets (time, latitude, longitude and variable). In addition, RCMES can now be viewed as an application of Apache OCW.

3.2.1 The Regional Climate Model Evaluation System

RCMES consists of two components: the Regional Climate Model Evaluation Database (RCMED) and the Regional Climate Model Evaluation Toolkit (RCMET) (Figure 3.4).

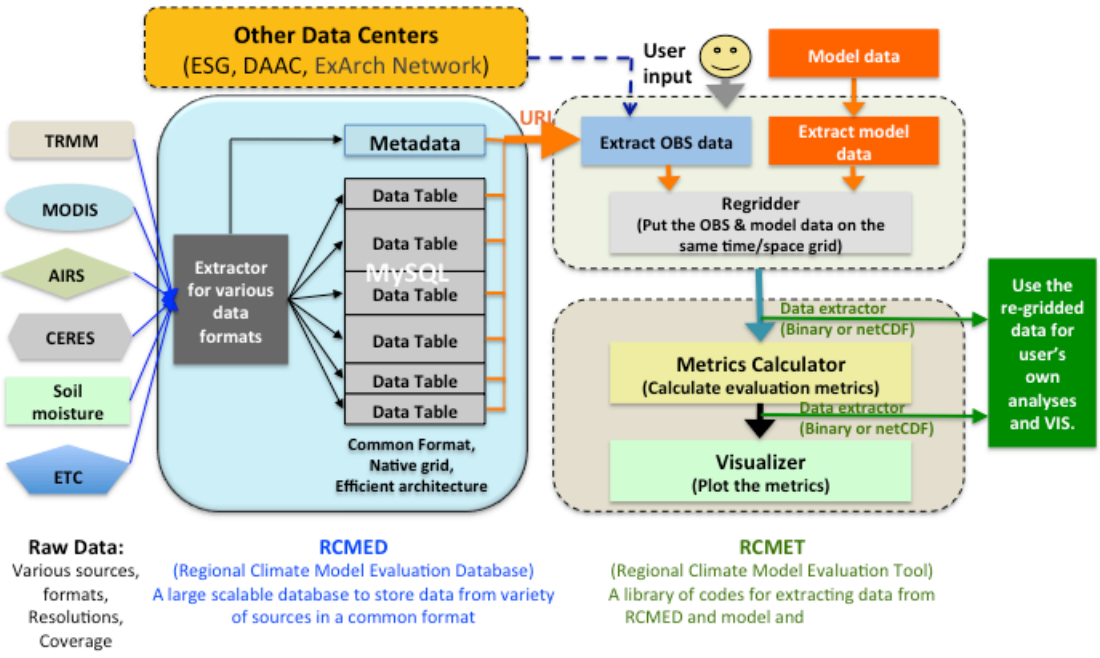


Figure 3.4. Schematic of RCMES. Adapted from Mattmann *et al*. (2013).

RCMES was designed for a workstation / PC environment, and on that account, the RCMET data access component quite cleverly stores data accessed from RCMED locally on the user's system, in a compressed format, based on the assumption that the user will access these data constantly for any respective studies (Hart *et al*., 2011). In that manner, time taken to access the database is significantly reduced, especially for those instances where Internet speeds are nominal. Additionally, this feature allows for the offline usage of RCMES. RCMED is a collection of PostgreSQL databases inclusive of NASA's remotely-sensing data such as TRMM, reanalysis data, and other observations, such as University of East Anglia Climate Research Unit datasets (CRU), that facilitates data classification, extraction and homogeneity, irrespective of the original data format. Examples include NETCDF 3/4, GRIB, and HDF 4/5. RCMED physically resides at JPL, and it easily accessible via RCMES (Crichton *et al.*, 2012). Ahead of an evaluation, users may be required to load their data into RCMED. RCMET provides the capabilities of regridding, calculating evaluation metrics such as root-mean-squared (RMS), bias and correlation coefficients, and visualizations of the data. RCMET's flexibility allows users to (1) change the workflow of an end-to-end evaluation, (2) supplement with their metrics, and (3) extract data during the evaluation (Mattmann *et al.*, 2013).

Integrating this research into RCMES would allow for search-and-clip functionality based on relationships between variables over time from one or more datasets to be added. Integrating in RCMED side will allow for clipping of the dataset in the database when it is requested, before transfer, which is advantageous for two reasons. Firstly, it will allow for services in RCMED to be utilized in the execution of the actual search-and-clip function, and secondly, it will allow for a smaller dataset to be transferred, thus reducing the time it takes to acquire a dataset. Adding the search and functionality into RCMET will allow its functionality to be used locally on the user's

system with user datasets (possibly not ingested into RCMED), and then request those clipped regions from RCMED in other datasets. The scope of this research considers the integration on the RCMET side.

3.2.2 The Apache Open Climate Workbench

The Apache Open Climate Workbench (Apache OCW) is an open-sourced Python-based climate evaluation toolkit that can be found within the Apache Software Foundation (ASF). Apache OCW is an ASF Top-Level Project, which is an endorsement statement from the ASF that signifies the project meets ASF's best practices. It is being used (and contributed to) by various organizations and universities including the National Aeronautics and Space Administration (NASA), the Coordinated Regional Downscaling Experiment (CORDEX) community, the Indian Institute for Tropical Meteorology, and the University of Cape Town, South Africa. The Apache OCW toolkit accesses observation data and model outputs from various organizations such as NASA, National Oceanic and Atmospheric Administration (NOAA), and the Earth System Grid Federation (ESGF).

Apache OCW was born out of the refactoring of RCMES code and donating that code to ASF. Through the refactoring process, the RCMES code was reconstructed (without changing the high-level functionality) to improve aspects of software development such as code readability, reducing complexity of functions by breaking them down to basic functionality and reducing dependency, and improving the code maintainability and extensibility. Apache OCW maintains the high-level functionality of RCMES, but allows the user greater flexibility in creating individual project workflow. From the refactoring process, Apache OCW contains the basic representation of RCMES – the "nuts and bolts" – in a reusable fashion for developers. Basically, Apache OCW allows users to build other applications that require the underlying

functionality of data extraction, manipulation, calculations, and visualization, and RCMES is an example of such an application. Furthermore, as Apache OCW is an open-sourced ASF project, users can contribute methods that are applicable to the overarching objective of the project. Some examples include creating various interfaces for performing evaluations such as a user interface (UI) and command line interface, a plotting application programming interface (API) and provision of methods to handle datasets that may be local to the user.

Given the relationship between Apache OCW and RCMES, it follows that overarching functionality between RCMES as a database and a toolkit, and Apache OCW are quite similar, but the distribution of the actual code (the software architecture) would vary.

The source code structure provides information as to how the software components within a software application should interact with each other. The Apache Open Climate Workbench (OCW) source code structure defines the set of files contain the classes or objects and the associated class methods, as well as their interactions in a quick and simple readable format (Figure 3.5). This facilitates users being able to quickly determine what is available from the project, and how to access or use that functionality.

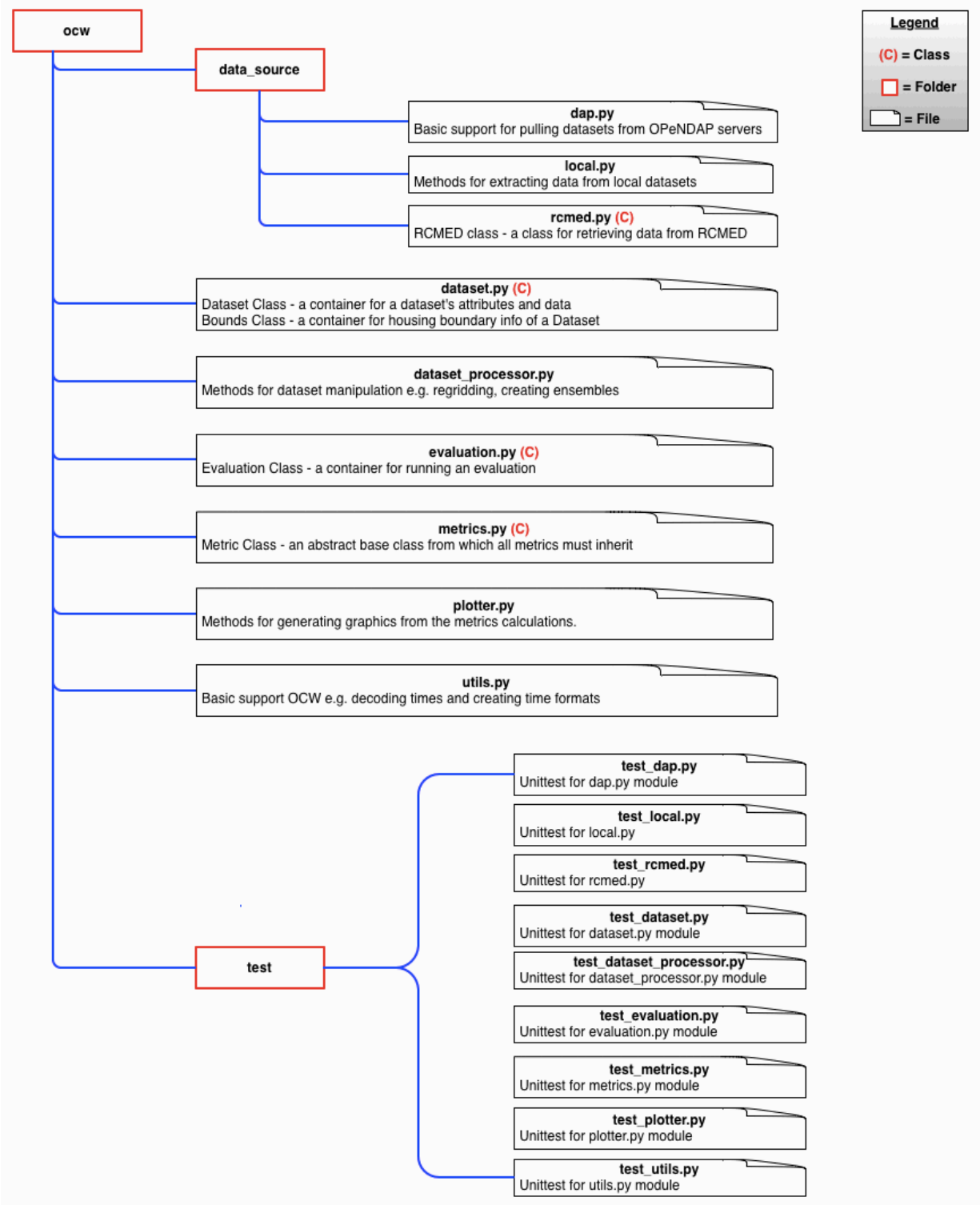


Figure 3.5. The Apache OCW Source Code Structure. [7]

[7] http://climate.apache.org/index.html

The Apache OCW provides methods on data extraction in the local.py file. Methods from this file are used in this project for extracting the satellite data into a common format. The dataset-processor.py file contains methods on regridding that are leveraged in this study. The metrics.py file provides methods on calculating metrics such as time and space averages, and standard deviations. Further, the plotter.py file contains methods for visualizing the metrics calculated based on the Python matplotlib library (Hunter, 2007). Methods from these two sources are also used in this study. Based on the API construct, it is expected the code from this project can be added to the Apache OCW through contributions of a new file that would sit in the main ocw folder (a file similar to evaluation.py). Additionally, metrics and visualizations developed in this study can be contributed to the metrics.py and plotter.py method collections respectively.

CHAPTER 4. A FULLY AUTOMATED SPATIO-TEMPORAL SEARCH FOR MCCs IN SATELLITE DATA

The framework for an automated tool to identify mesoscale convective systems (MCCs) addresses the limitations related to human intervention and handling voluminous data. From Chapter 2 it was observed that human intervention is paramount in past studies of MCCs for (1) identifying systems with complex evolutions (2) identifying precipitation characteristics. Removing this human intervention involves creating a method that addresses the aforementioned limitation. The method proposed here is based on the workflow of existing methods, but also incorporates a method in mathematics and computer science – graph theory – that reduces the demand on computation resources, amongst other advantages.

4.1 Graph Theory in the Atmospheric Sciences

In discrete mathematics and computer science, graph theory is used to model the relationships between objects (Trudeau, 1993). Graph theory refers to the study of these graphs and networks, where graphs are considered as objects, or more generally, a data object. Graphs naturally model the relationships and process dynamics in physical, biological and social systems, as well as naturally represent networks of communication and data organization. Graph theory involves determining the characteristics of graphs, such as their interconnectivity and traversal, as well as ways in which their structure (nodes and / or vertices) can be combined and arranged. Historically, graph theory has been heavily utilized for problems that require mapping functionality, like determining the best (shortest path and lowest cost) route for air travels on travel websites, timetabling / scheduling, and by telecommunication companies for development and maintenance of their network.

Graph theory, though not commonly implemented in the atmospheric sciences, has been used for the predicting (nowcasting and forecasting) of thunderstorms. Implementation of graph theory in the atmospheric sciences involves considering the relationship between atmospheric variables at a given time, or the spatial-temporal analysis of cloud volumes. In the first case, the vertices or nodes of the graph are considered to be the measured quantities in the atmosphere related to thunderstorm development, at a given time and atmospheric level, whereas the edges are the correlation between these variables. Chaudhuri and Middey (2009) and Chaudhuri and Middey (2011) illustrated this type of application. In the second method, combinatorial optimization of graph theory is implemented. The nodes of the graph represent a unique cloud mass at a given time, whereas the edges indicate correlation between cloud masses over time, as illustrated by Dixon and Wiener (1993) and Mukherjee and Acton (2002). In general, implementing graph theory in the domain of the atmospheric sciences appears to be innately applicable to the complexity, non-linearity and inherent chaos of the atmospheric system.

4.2 The "Grab 'Em, Tag 'Em, Graph 'Em" Algorithm for Tracking Mesoscale Convective Complexes

The algorithm developed for this study is coined to as the "Grab 'em, Tag 'em, Graph 'em" (GTG) algorithm. The GTG algorithm is based on the science of mesoscale convective system (MCS) identification, as implemented in previous studies but utilized graph theory. The graph theory used involves mostly using the graphs as an object along with some graph methods such as Dijkstra's shortest-path search (Dijkstra, 1959).

The "Grab 'em, Tag 'em, Graph 'em" (GTG) algorithm, like most automated (and semi-automated) algorithms for MCS identification, comprises of two main steps: (1) the cloud masking or cloud detection stage, and (2) the cloud tracking or evolution stage. The cloud

detection stage (as explained in Section 2.2.3.2) utilizes infrared (IR) satellite data and depends on brightness temperature (T_B) and area thresholds to identify regions of interest that could develop into various MCSs. These regions are referred to as cloud elements (CEs). The GTG algorithm uses a temperature threshold of 241 K for CE identification such that values warmer than the threshold are discarded and those colder are equal to it are maintained for analysis. This temperature is the criteria used for Maddox (1980b) outer shield, and is maintained here because it has been shown that warmer temperatures are associated with a precipitating shield in well-organized mesoscale systems (Adler & Negri, 1988). The GTG algorithm also uses an area cutoff of 2,400 km^2 in conjunction with the T_B threshold. This means that all contiguous areas where the T_B is less than or equal to 241 K are considered as area of interest to develop into an MCS. A further consideration, mirroring the scientific premise of Bouniol *et al.*, (2010), considers areas smaller than 2,400 km^2 where the T_B range within the potential identified contiguous area is at least 10 K. In this algorithm, each image of data is referred to as a frame (F). Each frame is a function of latitude, longitude and brightness temperature (T_B).

$$F_t = \{lat, lon, T_B\}$$

Within each frame, a number of cloud elements (CEs) exists. The properties of a CE in the region of interest in this study, as illustrated by Goyens, Lauwaet, Schröder, Demuzere and Van Lipzig (2011), are:

1. Brightness temperature, $T_B \leq 241$ K
2. Area, $A \geq 2400km^2$ or $A < 2400km^2$ and (T_B minimum in the area/ T_B maximum in the area) >= T_B range, where T_B range = 10K

$$CE[lat, lon, T_B]_t$$

$$\in \{F_t[lat, lon, T_B] \mid (F_t[lat, lon, T_B] \leq 233K \ \wedge ((F_t[lat, lon, T_B] \geq A\ )$$

$$\vee ((F_t[lat, lon, T_B] \leq A\ ) \wedge (\frac{F_t\left[lat, lon, T_{B\_min}\right]}{F_t\left[lat, lon, T_{B\_max}\right]} \geq T_{B\_range})\ ))\}$$

And,

$$F_t = CE[lat, lon, T_B]_{t,n} \in \{n \mid CE[lat, lon, T_B]_t\}$$

In the "Grab 'em, Tag 'em, Graph 'em" (GTG) graph implementation, the cloud elements (CEs) are the graph nodes (the "Grab 'em" part of the GTG). Furthermore, the process of identifying CEs requires them to be given unique identifiers ("Tag 'em" part of the GTG). Using graph theory nomenclature, during the cloud detection stage, a directed graph, *G*, is created where the graph nodes, *V(G),* are the CEs for each frame.

The second part of the automated (and semi-automated) algorithm for MCS identification involves tracking or evolving the CEs. The GTG algorithm completes this step simultaneously with the cloud detection stage. It is recognized that although each CE exists at a discrete time, there can be correlation amongst them over time that produces mesoscale convective complexes (MCSs). The GTG algorithm implements a combination of the common area-overlapping and maximum spatial correlation methods, as summarized in Section 2.2.3.2. From the literature, it was found that in the area-overlapping method, CEs that evolve into large-scaled MCSs are considered to be correlated if there is greater than 50 percent, or in excess of 10,000 km^2, area spatial overlap between successive images. In the maximum spatial overlap method, the percentage overlap varies according to the temporal resolution of the data – for example, 95 percent overlap for a maximum two-hour time difference (Goyens *et al.*, 2011).

In the GTG algorithm, CEs between frames are correlated via the percentage overlap and / or area-overlap between CEs of consecutive frames, and in the GTG graph implementation, are

represented by the graph edges. The edges of the graph, *E(G)*, are weighted according to percentage overlap between the geo-spatial parameters between cloud elements in frames F_t and $F_{t+\Delta t}$ and are directed from a parent node in F_t to a child node in $F_{t+\Delta t}$. The minimum weight of an edge occurs where the percentage overlap between the two CEs is at least 95 percent, whereas the maximum weight occurs for a 90 percent overlap, or in excess of 10,000 km^2 (Arnaud, Desbois & Maizi, 1992; Williams & Houze, Jr., 1987). Thus, a weighting function for the edges, $W_{lat,lon}$ is defined such that:

$W_{lat,lon} = (F_t CE_n \cap F_{t+\Delta t} CE_n)$(min=95% area overlap, max=90% area overlap or area overlap ≥ 10 000 km^2).

The final collection of cloud elements and edges is referred to as a cloud cluster (CC). A CC can have a minimum duration of two consecutive frames, and maximum of *m* consecutive frames, where *m* is number of frames being considered. As such,

$$CC_n = CE_{n,t} + CE_{n,(t+\Delta t)} + CE_{n,(t+2\Delta t)} + \cdots + CE_{n,(t+m\Delta t)}$$

It is noteworthy that cloud clusters (CCs) do not necessarily involve only one cloud element (CE) from a given frame. From observation as a CC evolves, CEs from different frames can participate in merging, splitting, growth, decay, or maintenance. At this point, the directed graph depicts all the CCs or MCSs found within the time period analyzed, where a CC is a subgraph of the directed graph, G. To identify MCSs, each subgraph would have to be searched to determine if it meets the criteria. A (deepest) Dijkstra shortest-path search is implemented on each subgraph to determine if the MCSs criterion and the MCC criterion for West Africa are met. By implementing the Dijkstra shortest-path search in a weighted directed graph, it is assumed that if merging and splitting occurs within a CC, there must be the minimum overlap between consecutive CEs.

Once the graph has been built, node information can also be used to perform search (and clip) functions in other datasets. This creates another object for which comparisons within other datasets can be conducted / made.

4.2.1 Implementing the "Grab 'Em, Tag 'Em, Graph 'Em" Method

The "Grab 'em, Tag 'em, Graph 'em" (GTG) implementation uses the numerical and scientific libraries in Python i.e. NumPy and SciPy (Oliphant, 2007). The graph theory is implemented using the Networkx Package (Hagberg, Schult & Swart, 2008), and visualizing is supported through matplotlib library (Hunter, 2007). Existing functionality in the Apache Open Climate Workbench project that reads datasets into five dimensional Numpy arrays (time, latitude, longitude, altitude, variable) is also utilized. In the GTG implementation all the datasets are voided of their original data format characteristics and are read into NumPy four-dimensional arrays (time, latitude, longitude, variable). As such, in theory, any four-dimensional dataset that can be represented in that format can implement the GTG algorithm.

Figure 4.1 illustrates the general workflow of GTG where the parts of the code for user development, such as metrics calculation and visualization, are highlighted, as these will vary according to user cases. The algorithm implementation is currently standalone, as it does not require database access during execution, and the GTG algorithm can work with only IR satellite data, or with other dataset(s).

4.2.1.1 The Cloud Detection and Graph Creation Implementation

The cloud detection part of the "Grab 'em, Tag 'em, Graph 'em" (GTG) algorithm is implemented using functions from numerical and scientific Python libraries. Specifically, the SciPy package provides a suite of libraries for multi-dimensional array image processing (scipy.ndimage). The segmentation in the ndimage package allows for separating objects of

interest from a background using probably intensity thresholding. The connectivity of the objects is defined by a structuring element instance. The information retrieved from the segmentation process is stored in directed graph, implemented using the Networkx package. NetCDF files are generated during this stage using the netCDF4 libraries, so the data will be available outside of the running of the application, allowing for further analysis and / or visualization (as illustrated in Figure 4.1).

Figure 4.2 provides the algorithm of this integral part of the GTG algorithm. In the general workflow of the GTG, as illustrated in Figure 4.1, this step referred to as “findCloudElements” occurs after the original infrared datasets have been read and extracted into a NumPy array.

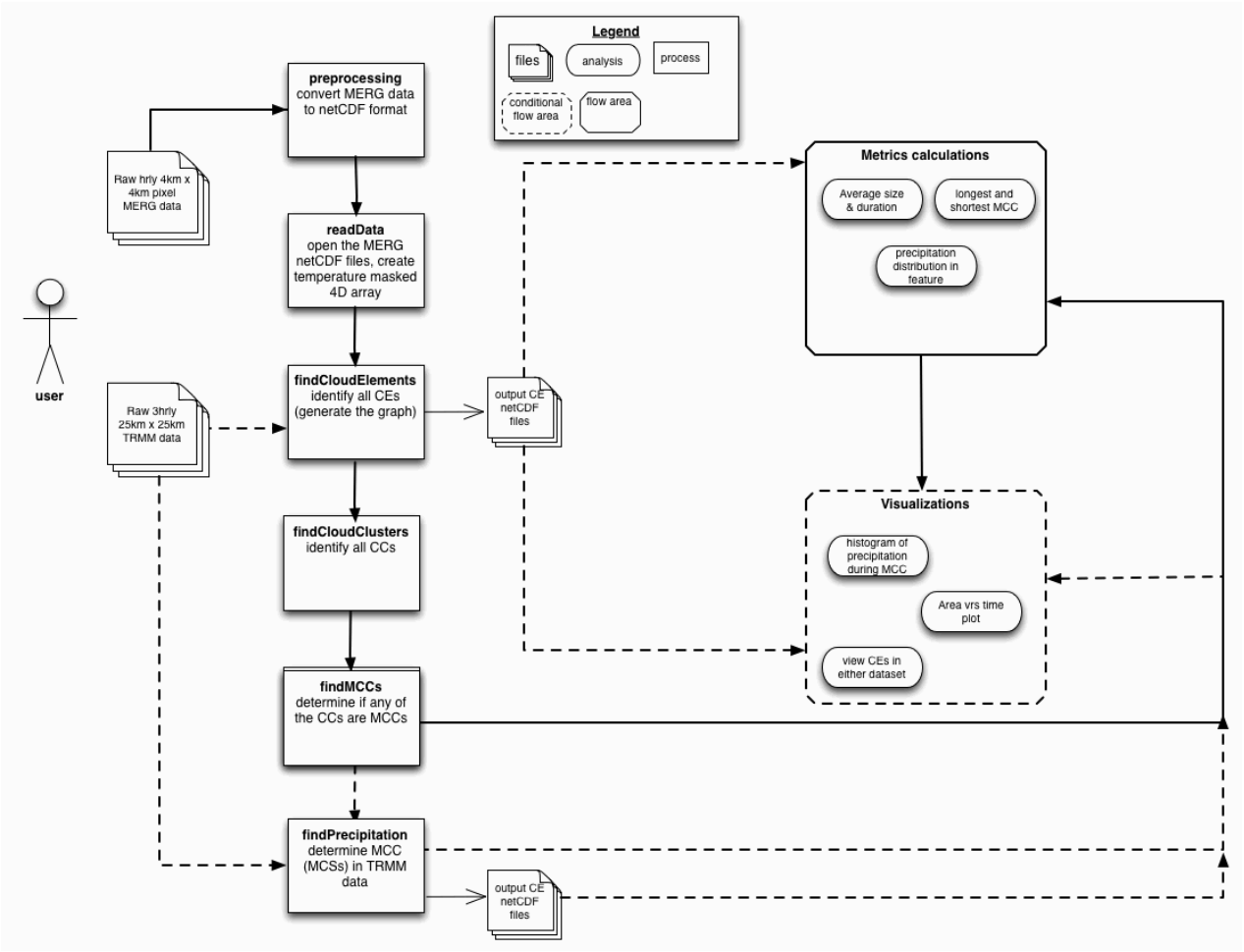


Figure 4.1. The general workflow of the "Grab 'em, Tag 'em, Graph 'em" algorithm.

Purpose:: Determines the cloud elements and their evolution in a series of frames from different datasets

Input:: *sat_img*: masked numpy array in (time,lat,lon,T_bb) representing the IR data.
TRMMdirName (optional): a string representing the file path to the local TRMM datafiles

Output:: *CLOUD_ELEMENT_GRAPH*: a directed Networkx graph of all CEs. The node of each graph is a dictionary – *cloudElementDict* – that contains unique information about the CE e.g. the unique CE identifier, CE time, area, etc.

BEGIN

foreach $F \in$ *sat_img* **do**
    determine the contiguous areas with T_B below 241 K using ndimage methods
    foreach *contiguous areas identified* $\in F$ **do**
        calculate the area
        if area criteria OR area and temperature range criteria
            contiguous area identified = CE
            create unique CE ID
            create a netCDF file with the CE ID as a filename
            store the lats, lons and T_B CE data in the netCDF file
            if *TRMMdirName*
                open the corresponding local TRMM datafile
                regrid the TRMM data and extract the relevant info
                store the lats, lons and precipitation data in netCDF file
            endif
            create *cloudElementDict* for this CE
            create *CLOUD_ELEMENT_GRAPH* node with *cloudElementDict*
            if this is not the first frame
                determine the edge existence and weighting
            endif
        endif
    end foreach contiguous area
end foreach frame

return *CLOUD_ELEMENT_GRAPH*

END

Figure 4.2. The algorithm for cloud detection and graph creation of the "Grab 'em, Tag 'em, Graph 'em" algorithm.

The Networkx package, along with matplotlib libraries, allow for visualization of this somewhat abstract process. Figure 4.3 provides insight into the cloud detection and graph creation implementation outputs for two locations. In one iteration through the satellite data (frames), all the areas of interest were immediately isolated and the complexity of evolving systems were captured. During that iteration the abstract concept of a node in the graphs, as depicted the uniquely labeled red dots in Figure 4.3, represents information about the CE. Such information includes the CE's unique identifier, its area, its centroid point, and the latitude, longitude and value points. Within the Networkx representation, this information is stored in a Python dictionary datatype.

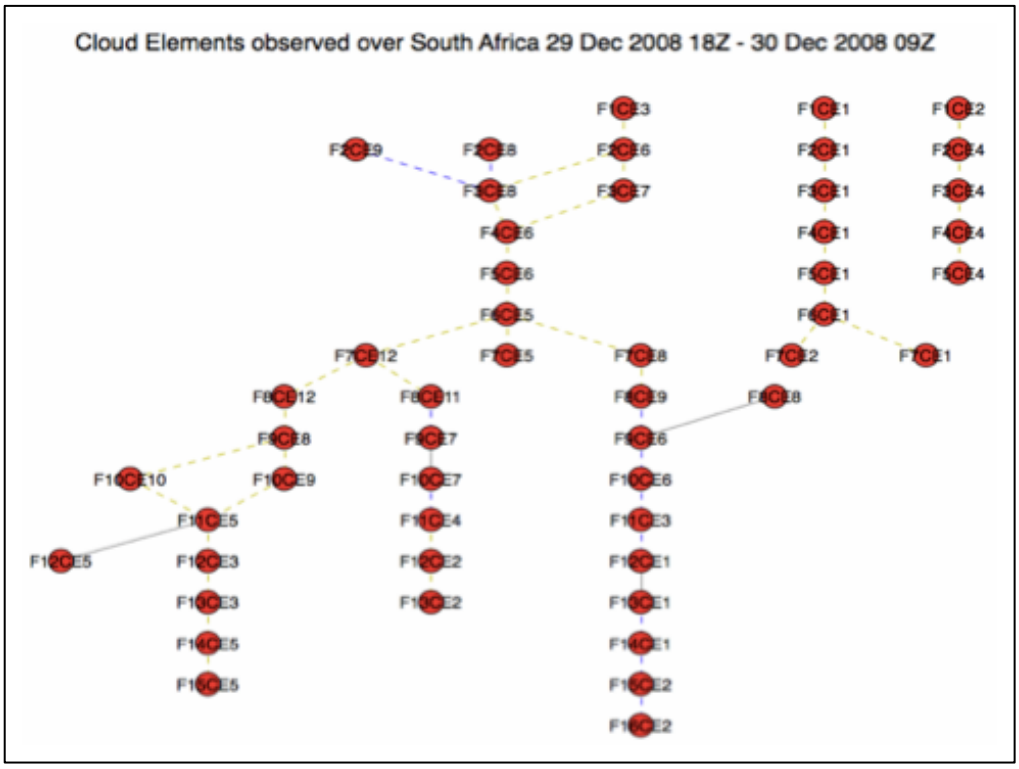


Figure 4.3. Cloud elements (CEs) observed after implementing the cloud detection and graph creation part of the "Grab 'em, Tag 'em, Graph 'em" algorithm. The red dots (graph nodes) represent the CEs identified. The text indicates the frame number and the CE number for that frame. The lines represent graph edges. Black lines indicate an area overlap ≥ 95 percent between CEs. Blue dashed lines indicate an area overlap between 95 and 90 percent. Yellow dashed lines indicate an area overlap ≥ 10,000 km^2.

4.2.1.2 The "Finding Cloud Clusters" Implementation

The "find cloud clusters" implementation executes a maximum depth and minimum path search on each node of the main tree to determine the cloud cluster (CC), if any, the node belongs to, thus pruning the original tree (Figure 4.4). This step occurs after finding all the cloud elements and constructing the graph, as illustrated in Figure 4.1. The minimum path search uses the Dijkstra shortest-path search (for weighted edges in this case). The Dijkstra shortest-path is an algorithm that identifies the minimum cost (the edges define the cost), or the shortest path to

traverse from one node of the graph to another. A common practical application of Dijkstra's shortest-path algorithm is finding the shortest route between one city and other cities, as would be commonly used by express mailing services. By adding the longest / deepest path criteria on to the Dijkstra's shortest-path, the furthest evolution of a cloud element (CE) can be tracked. The graph functions used are part of the Networkx package.

```
Purpose:: Determines the CCs from the subgraphs in main graph i.e. prunes the graph
according to a maximum depth and minimum path
Input:: CLOUD_ELEMENT_GRAPH: a directed Networkx graph of the CEs with
weighted edges
Output:: PRUNED_GRAPH: a directed Networkx graph of all CCs from the original data
BEGIN
checkedNodes = []
foreach node ∈ CLOUD_ELEMENT_GRAPH do
        if node ∉ checkedNodes
                find the Dijkstra shortest path and longest length for node in
                CLOUD_ELEMENT_GRAPH
                foreach pathNode ∈ Dijkstra shortest path do
                        if PRUNED_GRAPH ∌ pathNode
                                add pathNode to PRUNED_GRAPH
                                add related edge info from CLOUD_ELEMENT_GRAPH
                        endif
                end foreach pathNode
                update checkedNodes with pathNode
        endif
end foreach node
return PRUNED_GRAPH
END
```

Figure 4.4. The algorithm for determining cloud clusters in the "Grab 'em, Tag 'em, Graph 'em" method.

Figure 4.5 provides a visualization of this process, where the location in Figure 4.5 corresponds to those in Figure 4.3. Implementing the Dijkstra shortest- and longest-path searches on each node, immediately the complexity of the graph reduces, and some subgraphs from the cloud detection stage divide into separate subgraphs. To illustrate, the complex tree that ranges from F1CE3 to F16CE2 in Figure 4.3 reduces to two trees in Figure 4.5. In Figure 4.5, the first corresponding subgraph indicates that the tree originating at F1CE3 and ending at F16CE2 taking a simpler path, whereas the second subgraph that originated from the complex tree F1CE3 to F16CE2 in Figure 4.3 originated from F7CE12 and ended at F15CE5. Though not provided here, at this stage it is possible to visualize each CE on a geospatial map from the associated outputted netCDF files **(**Figure 4.1**).**

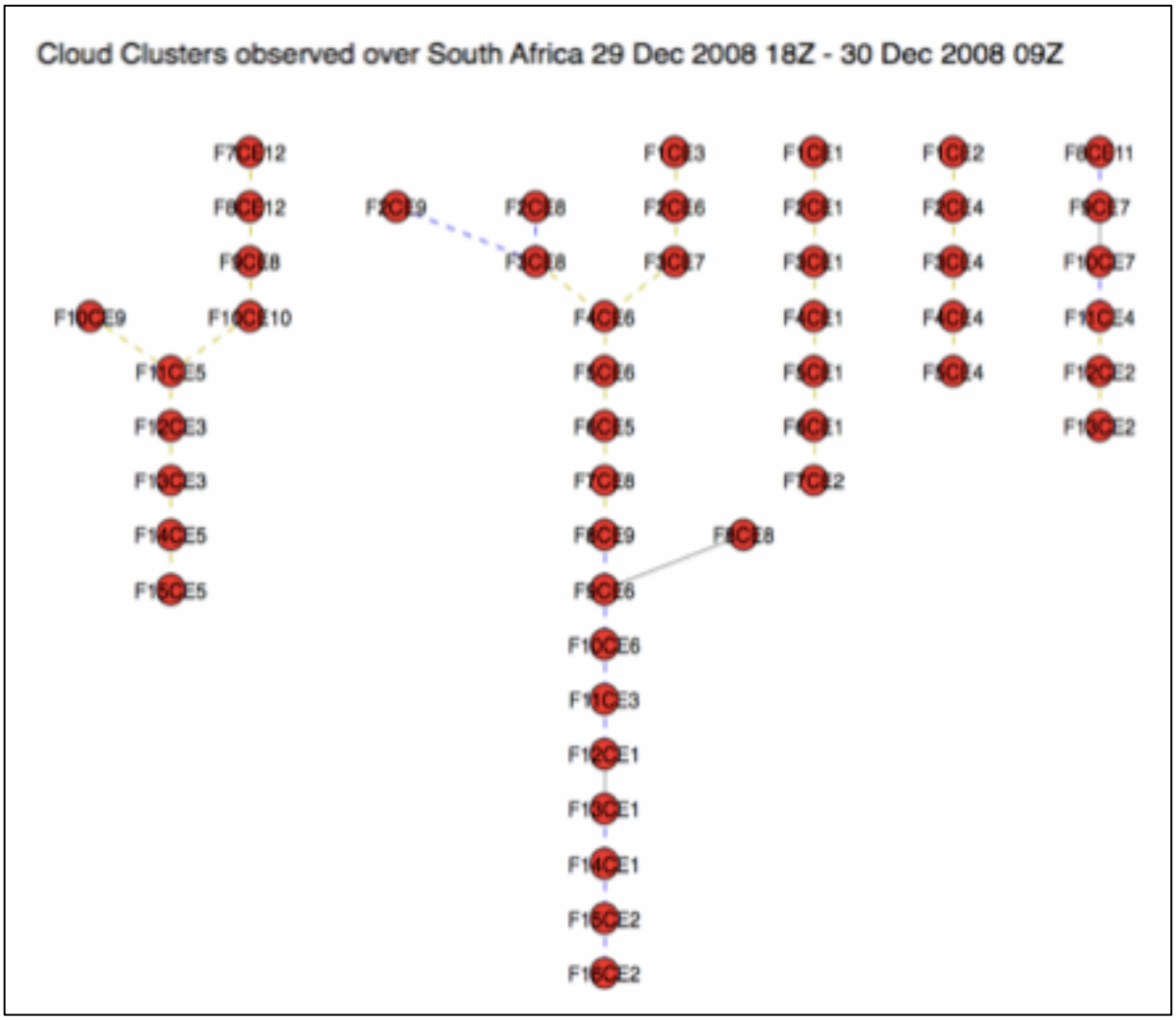


Figure 4.5. Cloud clusters observed after implementing the "find cloud clusters" part of the "Grab 'em, Tag 'em, Graph 'em" algorithm. The red dots (graph nodes) represent the CEs identified. The text indicates the frame number and the CE number for that frame. The lines represent graph edges. Black lines indicate an area overlap ≥ 95 percent between CEs. Blue dashed lines indicate an area overlap between 95 and 90 percent. Yellow dashed lines indicate an area overlap ≥ 10,000 km^2.

4.2.1.3 Mesoscale Convective Complexes Search Implementation

The purpose of the mesoscale convective complex (MCC) search implementation is to determine if any of the cloud clusters (CCs) (subgraphs / subtrees in *PRUNED_GRAPH*) meet the criteria (Table 2.2) for MCCs, as outlined by Laurent, D'Amato and Lebel, (1998), using the IR satellite data (MERG in this case). This step follows the "findCloudClusters" step as illustrated in **Figure 4.**. Notably, characteristics about the sub-graphs (alias sub-trees) that are created as a result of cloud cluster implementation will govern the graph methods used to facilitate the MCC searches. Figure 4.3 and Figure 4.5 demonstrate such characteristics are:

(1) The subgraphs created are directed, as the time increases from frame to frame. Furthermore, there is no direct connectivity between neighbors (nodes, or CEs in this case, within the same frame). Neighbors can however be connected through a node that splits thus creating children, or from merging nodes, where parents merge into one.

(2) The subgraphs are non-binary graphs as more than two nodes can be connected to one node from the previous frame.

(3) The complexity of the generated subgraphs varies. In some cases, there will be a 'simple' CC to search. A simple CC is defined as subgraph with (a) no merges or splits of CEs between frames; (b) only merging of CEs between frames; (c) only splitting of CEs between frames. In other cases, there will be complex subgraphs, where nodes both merge and split, though their frequency may be less than simple graphs.

(4) In general, the subgraphs will be very connected.

For the MCC search, a method to traverse the subtree to check each CE (node) and determine whether it meets the MCC criteria is required. Basic methods for searching directed non-binary trees include the depth-first search (DFS) and the breadth-first search (BFS). The DFS is an algorithm to traverse a tree in search of a node. The search starts at the root of the tree then explores as far as possible along a branch downwards as far as possible. It then backtracks to the root and explores another root in the same manner, until the desired node is found, or until the entire tree is traversed. The BFS is an algorithm to search and / or traverse a tree to find a node. The BFS starts at the root of the tree and laterally checks all nodes before exploring another level, until the desired node is found or the entire tree is traversed. There are limitations associated with either method. One such limitation is the DFS runs the risk of never terminating and being time inefficient, whilst the BFS is space inefficient, especially for large trees. In this study, a combination of a traversal method including DFS and BFS is required. A BFS traversal alone would discredit any new CEs (nodes) that were not connected directly to the head node, and thus this study merits the use of both DFS and BFS.

The depth-first iterative deepening (DFID) search algorithm is an optimization of the time-inefficient depth-first search (DFS) algorithm, and space-inefficient breadth-first search (BFS) algorithm (Korf, 1985). The search originates at the root node and a DFS search at a depth of one is performed, where all the nodes on that level are then checked in a BFS-like search. The tree traversal is continued in this fashion until the maximum depth of the tree, or the desired node, is reached. The main disadvantage of the DFID is the ’wasted computation’ performed prior to reaching the desired node. However, Korf, (1985) showed that the asymptotic growth of the running time of the DFID as a search algorithm is not affected by the ‘wasted computation’ time.

In this implementation, the idea is to reach the end of the tree and not to search for a node, as the Dijkstra's shortest-path was already implemented at the cloud cluster stage.

The DFID implements recursion to traverse (or search) a tree as far across on a level as possible in the tree. This includes searching complex structures during that breadth-wise path, before deepening. As such, complex branches of the 'main' tree can be searched before iterating within the depth of the tree. The algorithm for the modified DFID as implemented is illustrated in Figure 4.6, with examples of the traversal provided in Figure 4.7. Typically, the DFID implementation uses a stack datatype – a last-in, first-out (LIFO) datatype that allows for information to be added and removed from the same end of an entity. However, in this implementation, in order to account for searching the tree in the forward and backward direction, a modified stack datatype is used. Specifically, the stack behavior is modified in this implementation when searching backwards for parents of a merged node, where new information is added to frontend of the entity.

The complexity of a traversal can be expressed as a function of the branching factor in space, *b*; and the depth of the solution, *d*. The modified DFID performs a DFS to depth one. Then discards the nodes generated in the *d=1* search, start over and do a DFS for *d=2*, repeating this pattern until the end node is reached. Further, the modified DFID expands all nodes up to and below a given level (*d*) for each node identified in the DFS, in a recursive fashion. As such, the space complexity is $O(bd)$, and the time complexity $O(b^d)$.

Algorithm for traverseTree

Purpose:: To traverse a tree using a modified depth-first iterative deepening (DFID) search algorithm. Recursive implementation.

Input:: *subGraph*: a Networkx directed graph representing a CC

lengthOfsubGraph: an integer representing the length of the subgraph

node: a string representing the node currently being checked

stack: a list of strings representing a list of nodes in a stack functionality i.e. Last-In-First-Out (LIFO) for sorting the information from each visited node

checkedNodes: (optional) a list of strings representing the list of the nodes in the traversal

Output:: *checkedNodes*: a list of strings representing the list of the nodes in the traversal

BEGIN

if len(*checkedNodes*) = len(*subGraph*)

return *checkedNodes*

endif

if not *checkedNodes*

empty *stack*

update *checkedNodes* with *node*

endif

if the parents of *node* exists

add *node* to (front) of *stack*

endif

if children of *node* exists

add *node* to (end) *stack*

endif

foreach *eachNode* ∈ *stack*

if *eachNode* **not in** *checkedNodes*

update *checkedNodes* with *eachNode*

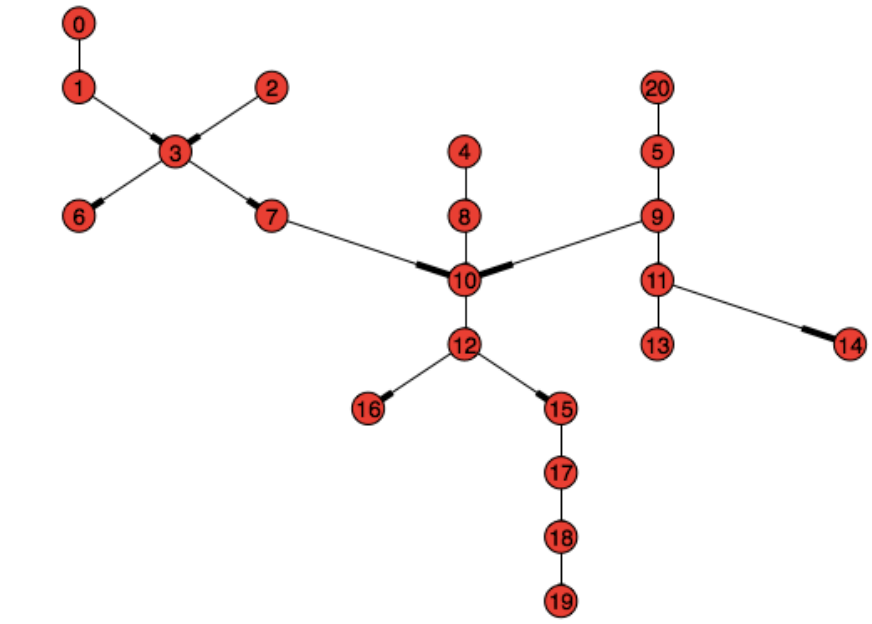
return traverseTree(*subgraph,eachNode,stack,checkedNodes*)

endif

end foreach eachNode

END

Figure 4.6. The algorithm for the modified depth first iterative deepening search implemented in this study.



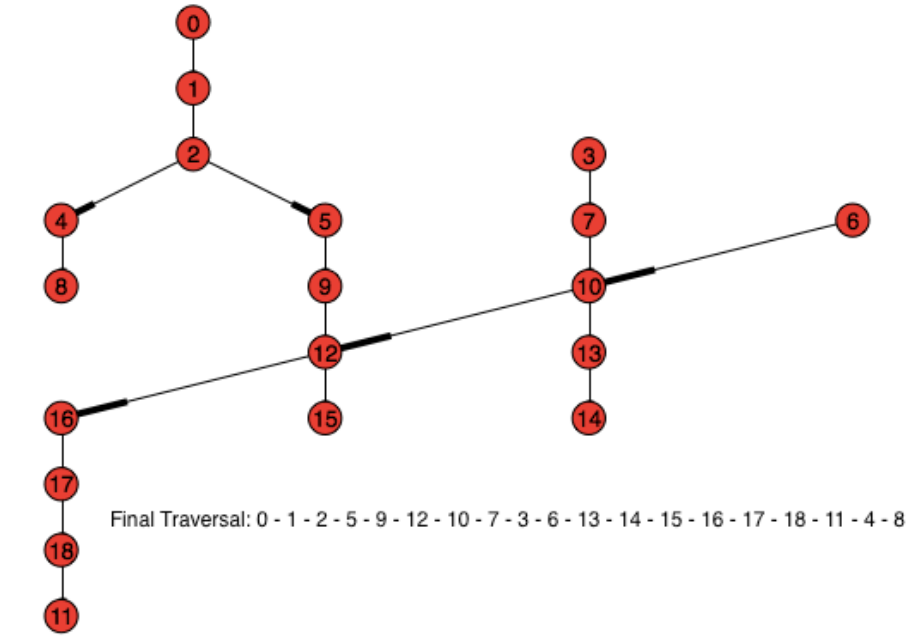


Figure 4.7. Examples of the results from the modified depth first iterative deepening search implemented in this study.

The mesoscale convective complex (MCC) in the "Grab 'em, Tag 'em, Graph 'em" (GTG) algorithm will be defined by consecutive cloud elements (CEs) that meet the area-temperature criteria, the duration criteria and the shape criteria. The MCC search not only identifies the feature, but also notes the various stages of development, from initiation, through maturity, to decay. Though the MCC search algorithm itself is very simple (Figure 4.8), it calls various utility functions to actually perform the checks. The MCC search implementation utilizes the Networkx libraries.

Purpose:: Determines if a subgraph (or part thereof) is a MCC according to Laurent *et al.* (1998) criteria
Input::: *PRUNED_GRAPH*: a directed Networkx graph of all CCs from the original data
Output:: *finalMCCList*: a list of dicts of nodes representing MCCs. The dictionary contains information about the MCC, its duration, the full MCC i.e. the developing and decaying CEs (nodes) that would not have meet the MCC criteria in a CC that contains a MCC. Also note that the nodes contain lat,lon data in a 4D array (time, lat,lon,T_B)
finalMCCList: a list of dicts of nodes representing all MCSs
BEGIN
foreach *subgraph* ∈ *PRUNED_GRAPH* **do**
treeTraversalList = a modified DFID traversal to determine the order of the nodes
check the *treeTraversalList* to determine if MCCs are a part of that tree
update *finalMCCList*
update *finalMCSList*
store relevant information in a text file
end foreach subgraph
return *finalMCCList, finalMCSList*
END

Figure 4.8. The algorithm for finding mesoscale convective complexes according to the Laurent, D'Amato and Lebel (1998) criteria in the "Grab 'em, Tag 'em, Graph 'em" method.

4.2.1.4 Tracking the Feature Using Tropical Rainfall Measuring Mission Data

Once the mesoscale convective current (MCC) (or the mesoscale convective system (MCS)) has been determined from the MCC search algorithm within the MERG data, the identified feature(s) can be extracted from other datasets, such as the Tropical Rainfall Measuring Mission (TRMM). This functionality in the "Grab 'em, Tag 'em, Graph 'em" (GTG) method can occur in one of two ways: either, (1) initially when the cloud elements were being identified, the location to the TRMM files (*TRMMdirName* in the "find cloud elements" algorithm in Figure 4.2) was provided, so that data can be extracted and stored in the node's dictionary, or (2) after the MCC has been identified in the MERG dataset, individual TRMM files (or other identified files) can be searched. Specifically considering TRMM data, the function examines how to find the precipitation rates and totals associated with the MCC feature (Figure 4.9).

The "find precipitation rates" algorithm requires regridding of the data so that the datasets can be compared. A bilinear spatial regridding is done to bring the lower resolution data to the higher resolution dataset, utilizing existing modules from the Apache OCW. The corresponding latitude and longitude locations are then extracted and stored for further metrics calculations and post-processing.

```
Purpose:: Determines the precipitation rates for MCSs found
Input:: CEdirName: a string representing the directory for the generated CE netCDF files
        TRMMdirName: a string representing the directory for the TRMM netCDF files
        MCSList: a list of strings representing the nodes that contribute to a MCS (MCC)
Output:: allCEnodesTRMMdata: a list of dictionaries of the TRMM data
         NB: also creates netCDF with TRMM data for each CE (for post processing)
BEGIN
foreach node ∈ MCSList do
        open the corresponding CE netCDF file that was generated during the cloud
        element detection stage
        determine the lat,lon limits of the CE
        open the corresponding TRMM file for the time (frame number)
        regrid (lat and lons) the TRMM data to the CE data
        extract the TRMM data within the lat,lon limits and store as a netCDF file
        store extracted TRMM data as a 4D array (time, lat, lon, precip_rate) i.e. same
        format at the CE data, in the dictionary
        calculate area precipitation data
        determine maximum and minimum precipitation rates within the CE
        update allCEnodesTRMMdata
end foreach node
return allCEnodesTRMMdata
END
```

Figure 4.9. The algorithm for finding precipitation rates and totals in the "Grab 'em, Tag 'em, Graph 'em" method.

At this stage of running the "Grab 'em, Tag 'em, Graph 'em" (GTG), there is ample data for post-processing and visualization.

CHAPTER 5. CASE STUDIES IMPLEMENTING THE "GRAB 'EM, TAG 'EM, GRAPH 'EM" METHOD

In this section, three case studies will be used to explore the accuracy of the "Grab 'em, Tag 'em, Graph 'em" (GTG) algorithm, where possible results will be compared with existing methods and existing values for various parameters. For all of the case studies, the area considered is within West Africa.

5.1 Case study 1: Tracking a Mesoscale Convective System in Niamey, Niger

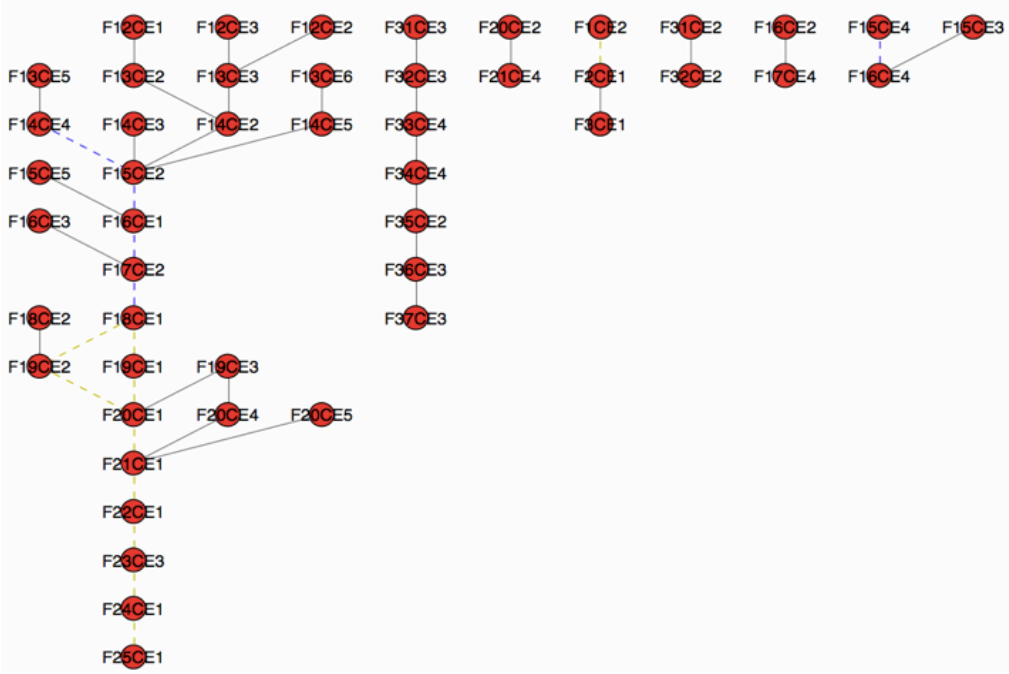
The primary objective of this case study is to demonstrate the accuracy of the "Grab 'em, Tag 'em, Graph 'em" (GTG) algorithm for the purposes of identifying a mesoscale convective system (MCS) in an infrared (IR) dataset, as compared to other methods. The secondary objective is to identify the precipitation characteristics associated with the feature. A MCS feature that occurred over Niamey, Niger (12°N–17°N, 8°E–8°W) is examined for the 36-hour period between 0000 UTC 11 Sep 2006 to 1200 UTC 12 Sep 2006. The 4 km MERG dataset resolution implies that a total of 60,280 spatial data points are considered in this domain, with 440 points in the x-direction and 137 data points in the y-direction. The run involves analyzing a total of 2,170,080 data points for each dataset. This case study is easily conducted on a personal portable computer.

This location and time period is considered because other methods specifically the Tracking of Organized Convection Algorithm through a 3-dimensional Segmentation (TOOCAN) algorithm, as illustrated by Fiolleau and Roca (2013), and the area-overlapping method, as outlined by Mathon and Laurent (2001), observed this feature that occurred during the African Monsoon Multidisciplinary Analysis (AMMA) project. The AMMA project is an international

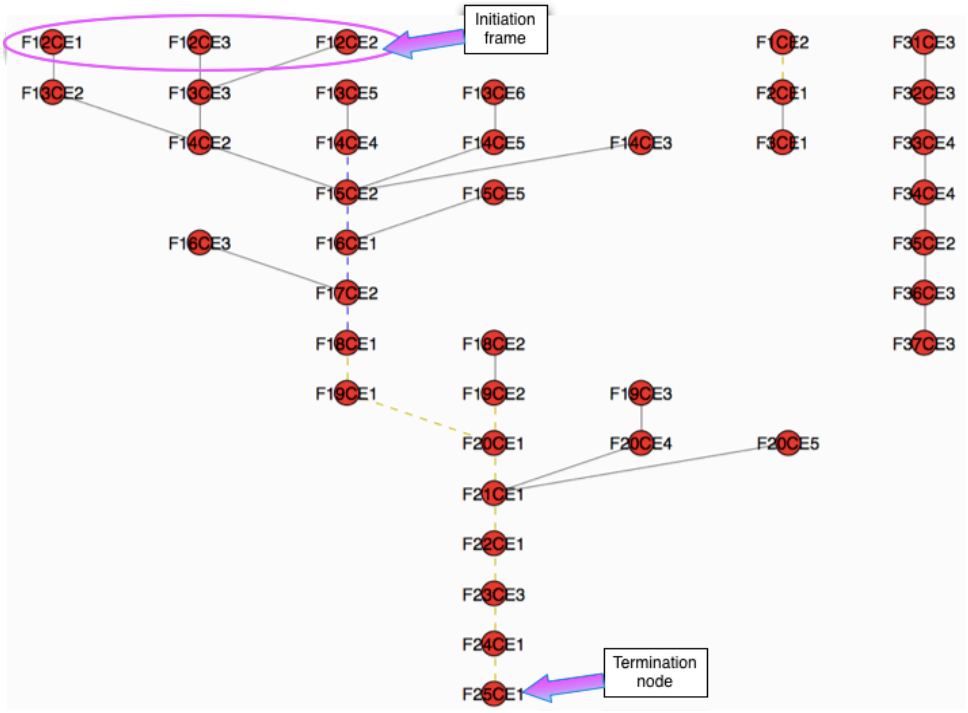
project geared at improving and increasing the scientific understanding of the West African Monsoon (WAM) with an emphasis of its variability on daily to interannual timescales (Redelsperger *et al.* 2006). During this project, measurements of MCSs over West Africa were made using a number of instruments including in situ radar, aircraft measurements and of course satellite data during summer 2006. The results of the area-overlapping method, as derived by Mathon and Laurent (2001) were presented in the study completed by Fiolleau and Roca (2013). Additionally, Fiolleau and Roca (2013) used infrared (IR) data from the METEOSAT second generation (MSG-1) in the 10.8 μm channels with 3 km and 15-minute resolutions and a 235 K brightness temperature (T_B) threshold was implemented. The reader is directed to Bouniol *et al.* (2010) for more information regarding the meteorology of the feature.

5.1.1 Comparing the Cloud Detection and Tracking of the "Grab 'em, Tag 'em, Graph 'em" Method with Other Methods

During the time period, the Grab 'em, Tag 'em, Graph 'em" (GTG) identified seven cloud clusters (CCs) of varying complexity and duration (ranging from two hours to 15 hours), as illustrated in Figure 5.1. Of the seven CCs, three were considered as mesoscale convective systems (MCSs). The MCS in question is the feature that appeared on 1100 UTC 11 Sep 2006 (node F12) and lasted until 0200 UTC 12 Sep 2006 (F25). This MCS originated from three cloud elements at 1100 UTC, specifically F12CE1, F12CE2 and F12CE3. Furthermore, two of these CEs - F12CE2 and F12CE3 - merged to form a CE at 1200 UTC (F13CE3), while F12CE1 evolved into F13CE2 at the time. In addition, at 1200 UTC, new CEs (F13CE5 and F13CE6) were formed. Splitting of CEs also occurred within this MCS. More specifically, at 1800 UTC, CE F19CE3 split, being observed as two CEs, F20CE1 and F20CE4, at 1900 UTC. In general, this MCS demonstrates a complex structure.



(a)



(b)

Figure 5.1. (a) Cloud clusters and (b) mesoscale convective systems identified over Niamey, Niger between 0000 UTC 11 Sep 2006 and 1000 UTC 12 Sep 2006 by the "Grab 'em, Tag 'em, Graph 'em" method.

Figure 5.2 shows the MCS in the original IR data, and the identified CEs from the GTG algorithm. Immediately, general features in the original IR dataset are identified. Furthermore, at some times, such as at 1700 UTC (F18), delineation of the large area identified in the original IR data is provided, as individual CEs are identified (F18CE1, F18CE2, F18CE3). From the original IR images, these GTG-identified CEs are areas of deep convection. After implementing the Dijkstra's shortest-path, as outlined by the CCs in Figure 5.1(b), the GTG identified a correlation between the CEs at 1700 UTC (F18) and 1800 UTC (F19) such that the smaller CE, F18CE1 is correlated with the larger part of the system (F19CE2) and F19CE1. This is shown as the dashed line in Figure 5.3. By 1900 UTC (F20) the two CEs from 1800 UTC are observed as one CE. This splitting and merging detail within the system has potential application to determining core locations of high precipiating rates in large-scale features.

Additional analysis and comparison of the identification of the CEs against the results of the TOOCAN algorithm and the area-overlapping method further validates the "Grab 'em, Tag 'em, Graph 'em" (GTG) methodology. The TOOCAN algorithm identified 32 MCSs, whereas the area-overlapping method identified five. The comparable definition of MCS between these two algorithms and the GTG is the cloud cluster, and the GTG identifies seven of these features. A limitation of the area-overlapping method is the creation and / or dissipation of MCSs by 'unnatural' splitting or merging of CEs. In this case study, the area-overlapping method initiated two 'unnatural' events. Neither the TOOCAN algorithm nor the GTG algorithm exhibits this behavior. Furthermore, the GTG identified decay (splitting) and regeneration (merging) stages in the morphology of the MCSs as the complexity of the graph.

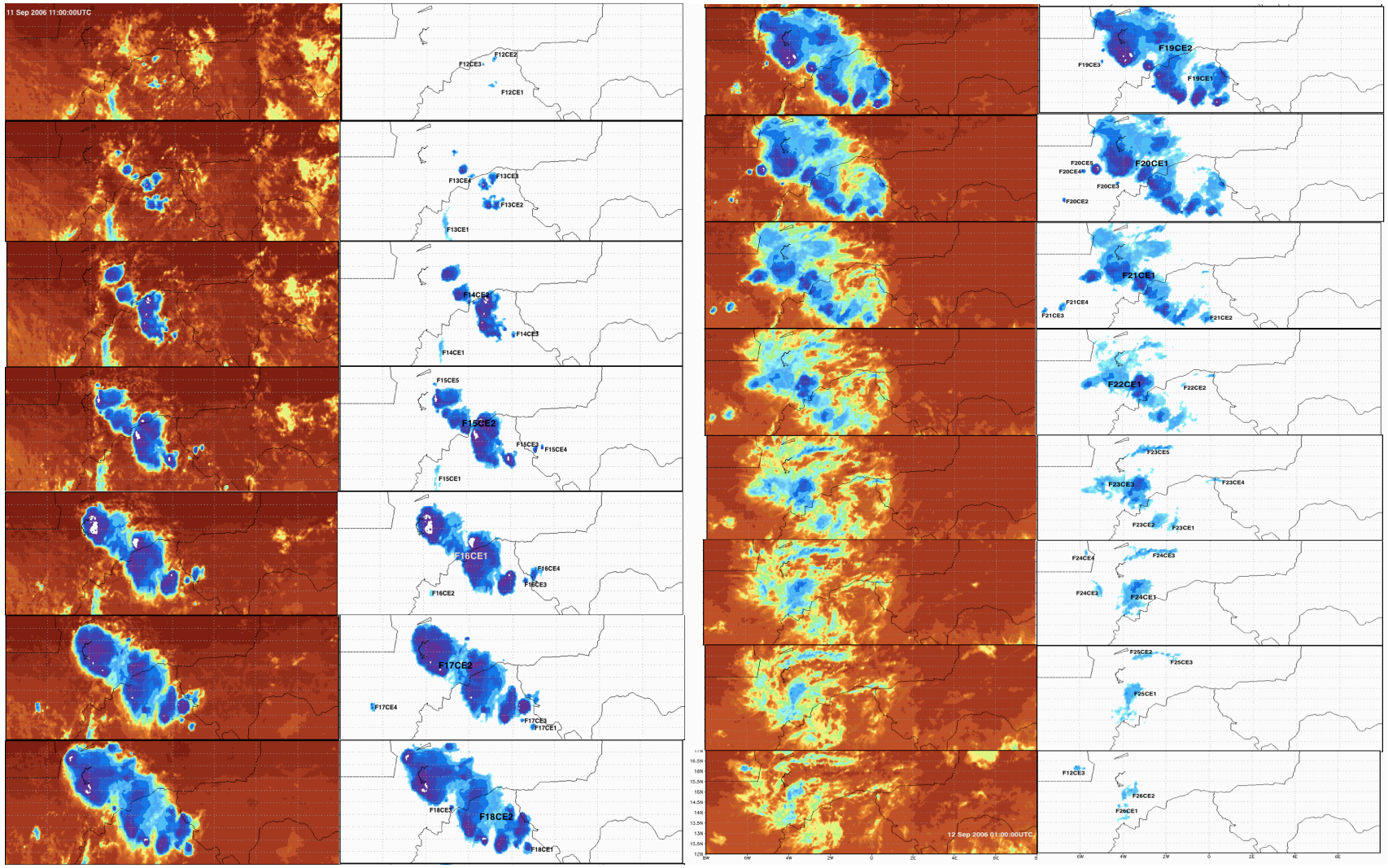


Figure 5.2. Cloud detection and tracking of the MCSs from 1100 UTC 11 Sep 2006 to 0200 UTC 12 Sep 2006 over Niamey, Niger. The first column illustrates the MERG images, the second column outlines the MCSs detected by the algorithm presented.

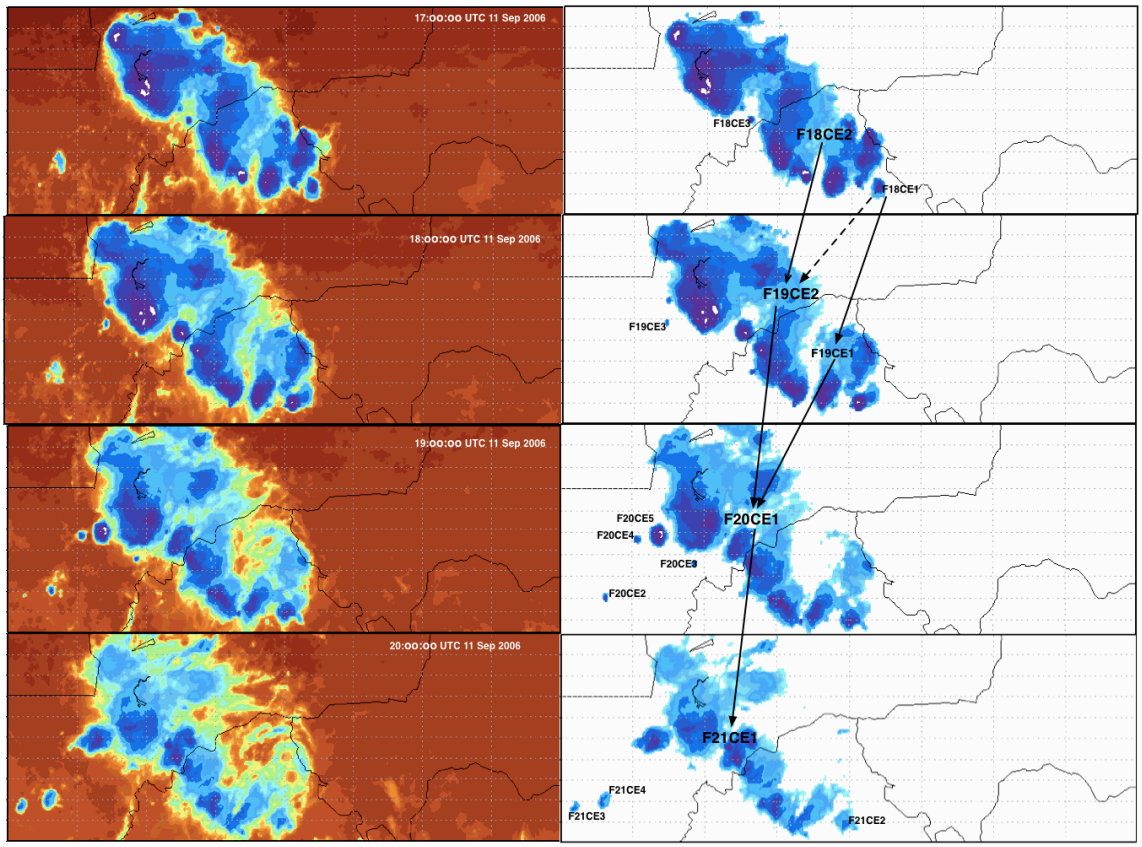


Figure 5.3. Cloud detection and tracking of the MCSs from 1700 UTC 11 Sep 2006 to 2100 UTC 11 Sep 2006 over Niamey, Niger. The first column illustrates the MERG images, the second column outlines the MCSs detected by the algorithm presented. The arrows indicate the connections between the CEs after the Dijkstra's shortest-path search (solid arrows). The dashed arrow indicates an edge that not a part of the shortest-path found.

The TOOCAN algorithm, which identifies the convective seeds, or areas of deepest convection, and iteratively generates the MCS from this radial point outwards, initiated the feature at 1200 UTC 11 Sep 2006, whereas the area-overlapping method initiated the feature at 1300 UTC 11 Sep 2006. The "Grab 'em, Tag 'em, Graph 'em" (GTG) initiated the feature at 1100 UTC 11 Sep 2006. Like the other methods, the GTG algorithm recognizes this feature as a large MCS, but does not identify it as a MCC.

The MCS reached its maximum extent around 2000 UTC with a maximum area of 177,392 km^2. This value is larger than the TOOCAN algorithm (79,000 km^2) and the overlapping method (162,000 km^2). These size differences are related to the temperature threshold used in the studies and the varying infrared datasets. The CE detection approach in the GTG captures the large-scale shape of the feature, including the warmer parts of the system, and in a non-iterative approach. This factor becomes increasingly important when the precipitation associated with the feature is being estimated.

In general, the "Grab 'em, Tag 'em, Graph 'em" (GTG) algorithm identified the same major mesoscale convective system (MCS) as the TOOCAN algorithm one hour earlier (at 1100 UTC 11 Sep 2006) and two hours earlier than the traditional overlapping method. Additionally, the GTG algorithm identified more MCSs and more independent features than the traditional overlapping method. The identification of convective seeds was not as efficient as the TOOCAN algorithm; nonetheless the identification of CEs surpassed that of the area-overlapping method, and the temperature data was maintained within each identified CE. The premise of the GTG method is fundamentally opposite to the TOOCAN algorithm, in that it seeks to delineate an entire cloud area associated with a feature as oppose to convective cores. Furthermore, the original temperature information is retained in this delineated area and stored in netCDF format for post-processing to meet individual research needs.

5.1.2 Using the "Grab 'em, Tag 'em, Graph 'em" Method to Determine Precipitation Characteristics

Neither the TOOCAN algorithm nor the area-overlapping method provides an automated method for determining precipitation characteristics associated with the feature. The "Grab 'em, Tag 'em, Graph 'em" (GTG) implementation of the characterization of the mesoscale convective

system (MCS) from the Tropical Rainfall Measuring Mission (TRMM) rainfall data indicates that, during the event, rainfall accumulations over the location ranged from 10 mm to 120 mm, although a small area, as compared to the entire precipitating area, actually received the larger totals (Figure 5.4).

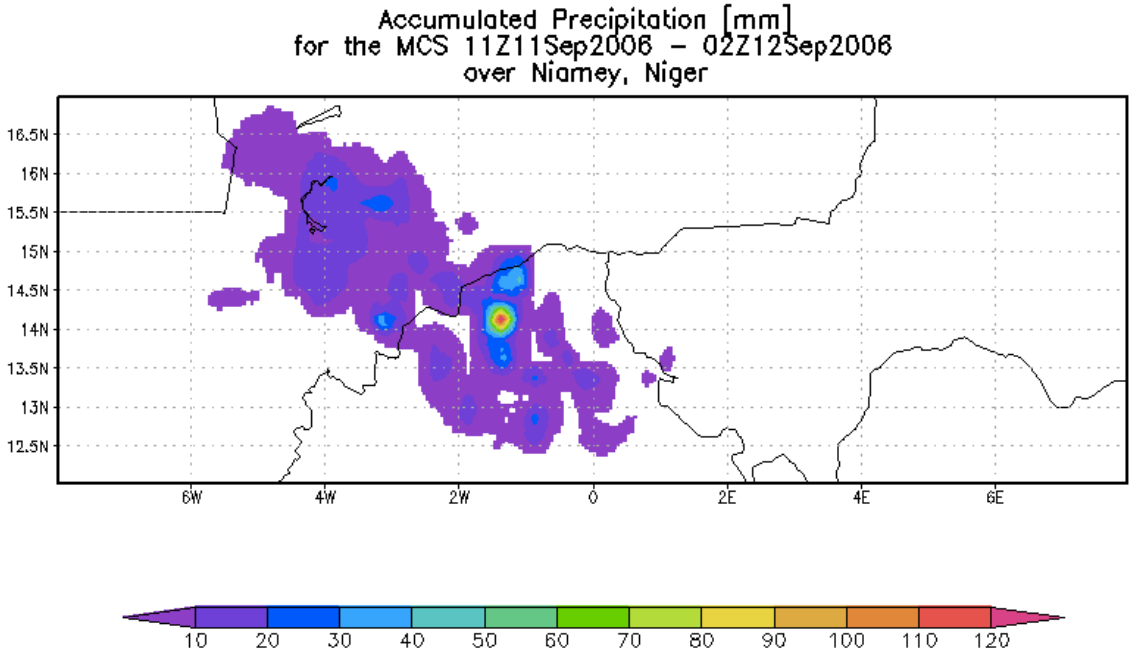


Figure 5.4. The accumulated precipitation in mm for the duration of the MCS over Niamey, Niger between 1100 UTC 11 Sep 2006 and 0200 UTC 12 Sep 2006.

Figure 5.5 provides an idea of when, where, and the rates of rainfall during the event. The initial stages of the MCS indicate that the cloud elements were mostly precipitating, as indicated by the large percentages. In the mature stage of the MCS, the precipitation area was greatly

reduced as compared to the feature. The feature was most active between 1600 UTC and 1800 UTC 11 Sep 2006 (Figure 5.6).

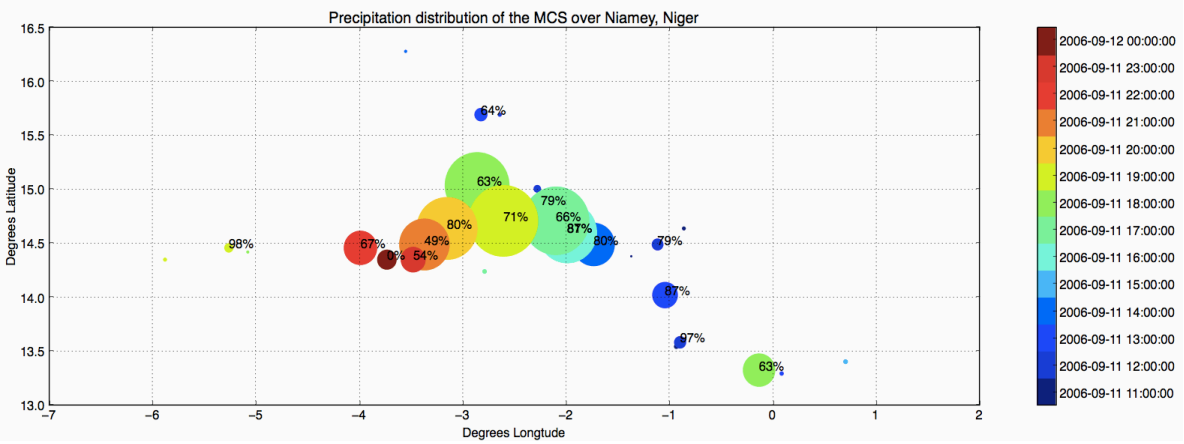


Figure 5.5. The spatial and temporal distribution of rainfall for cloud elements ≥ 2,400 km^2. The circles represent the relative area of each cloud element to the total area of the system. The percentage of the cloud element that was precipitating is illustrated.

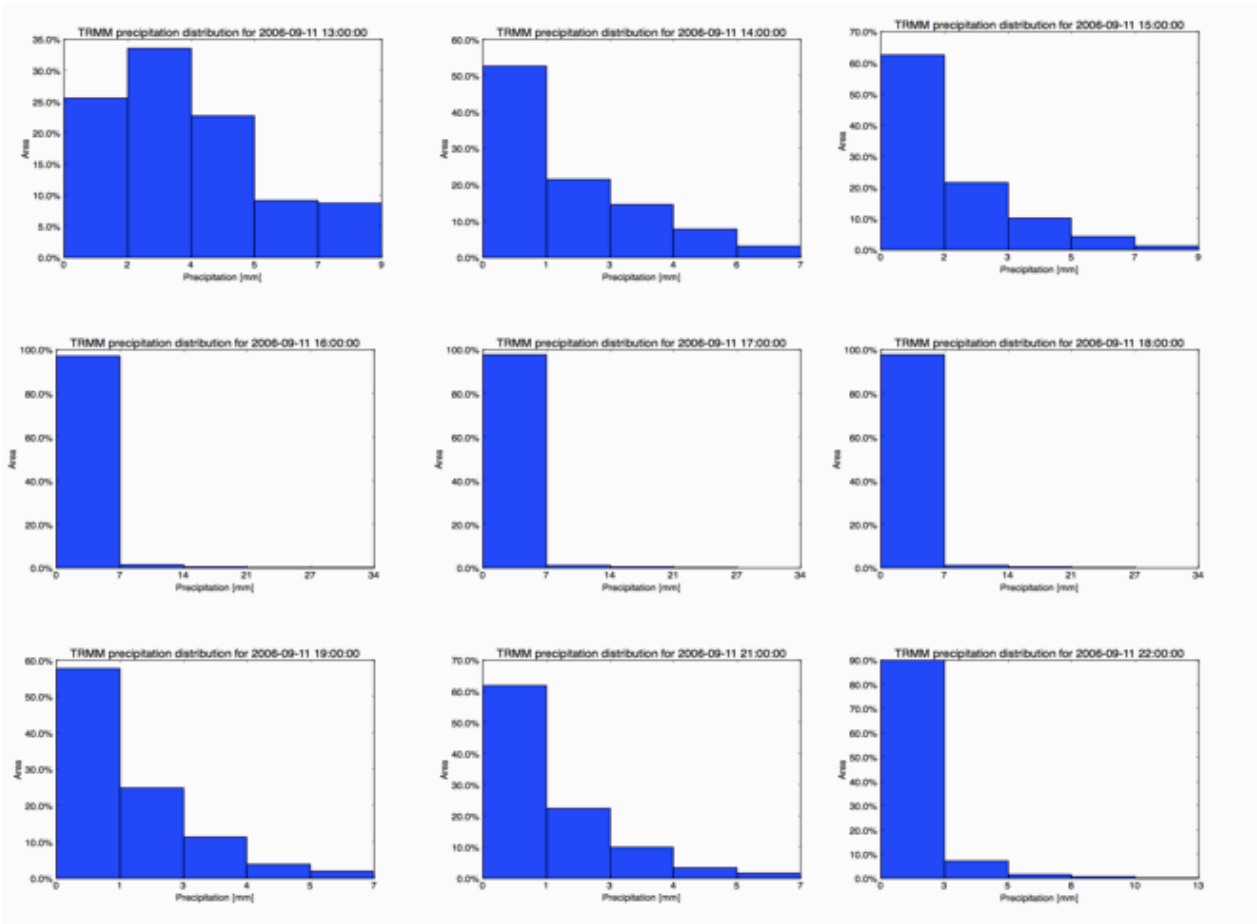


Figure 5.6. TRMM distribution of rainfall for selected times.

5.1.3 Summary of Case Study 1

The "Grab 'em, Tag 'em, Graph 'em" method demonstrated its graph implementation innately captures MCS morphology and its lifecycle. The GTG method accurately identified the mesoscale convective system feature in the IR dataset. The precipitation characteristics associated were also identified during the automated run, which provides a functionality no other existing method has. The single case event ran on a personal computer in 3.4 minutes.

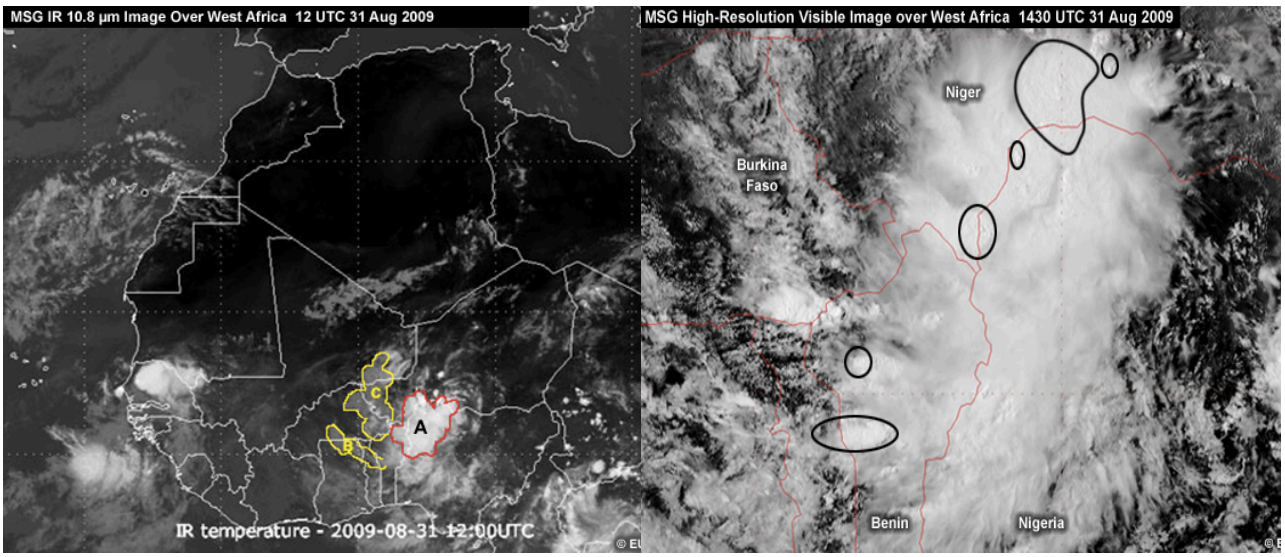
5.2 Case Study 2: Tracking a Mesoscale Convective Complex in Burkina Faso

The main objective of this case study is to illustrate the performance of the algorithm in identifying and characterizing a mesoscale convective complex (MCC) feature in an infrared and precipitation rate dataset. The African Easterly Wave (AEW) of 28 Aug 2009 to 5 Sep 2009 is a well-documented weather feature because of its socio-economic impact. A small mesoscale convective system (MCS) feature was initiated and rapidly grew to form an MCC (~240,000 km^2) by 1500 UTC 31 Aug 2009 (Galvin, 2010). The MCC feature brought heavy rain across the countries of West Africa, specifically Niger, Burkina Faso, southern Mali, northern Côte d'Ivoire, Guinea, Senegal, The Gambia and Guinea-Bissau between 31 Aug and 3 Sep 2009, where the intense rainfall lead to rivers breaking their banks and floods affecting 600,000 persons across these countries (Floods displace thousands, 2009). The feature received a lot of media attention, especially in Burkina Faso, where the daily rainfall total was 263 mm or approximately 22 percent of the annual rainfall (from gauge measurements), which lead to over more than 150,000 persons fleeing their homes (Burkina Faso 2009; West Africa floods, 2009).

The synoptic conditions were favorable for MCC development. These conditions included low-level wind convergence, mid-level trough, an area of outflow aloft, and favorable wind shear aloft. The reader is directed to the Flooding in West Africa (2011) Module on COMET for further details. The IR 10.8 micrometer image for 1200 UTC from the METEOSAT satellite (Figure 5.7(a)) indicates a large continuous and persistent cloud pattern with the coldest, most convective part to the northeast of Burkina Faso (A- the red outline area), and two large areas of warmer, lower (medium level) cloud ahead of the convective region (B & C – the yellow outlined areas). The high resolution VIS (HRV) image at 1430 UTC (1200 UTC is unavailable) indicates a well-defined MCS over Niger-Nigeria with towering anvil cloud tops from the most

convective parts of the feature casting shadows (outlined in black in Figure 5.7(b)). Note that the feature is very linear (as opposed to circular) and as such, it may not meet the shape criteria as outlined by Laurent *et al.* (1998). In practice, the manual identification of MCCs is generally not given a strict shape criterion because MCCs are usually embedded in other systems and/or sufficiently short-lived that this property is not clearly defined.

This location and time period are considered because a significant MCS event impacted the region during this time, which was subsequently covered by experts in meteorological studies such as the Flooding in West Africa (2011) Module on COMET and Galvin (2010). Thus, a comprehensive and agreed upon analysis that uses various datasets prepared by experts is available for the feature. The domain over Burkina Faso (5°N–19°N, 5°E–9°W) is examined for the 48-hour period between 0000 UTC 31 Aug 2009 and 2300 UTC 1 Sep 2009. The 4 km MERG dataset resolution implies that a total of 161,408 spatial data points are considered in this domain – 416 points in the x-direction and 388 data points in the y-direction. The run involves analyzing a total of 7,747,584 data points for each dataset. The case study took 7.6 minutes to run on a personal computer system. The feature is explored with a range for the shape criterion ($0.5 < \text{eccentricity} \leq 1.0$).



(a) (b)

Figure 5.7: (a) Annotated IR 10.8 micrometer imagery on 1200 UTC 31 Aug 2009 over West Africa. The MCS over Burkina Faso is considered here. A- the red outlined area represent the coldest part of the MCS and the B & C – the yellow outlined areas represent the warmer non-convective areas. (b) The annotated high resolution VIS satellite image for the same MCS at 1430 UTC. The black outlined areas indicate regions of deep convection.

5.2.1 Tracking the mesoscale convective complex with MERG data using "Grab 'em, Tag 'em, Graph 'em" Method

The "Grab 'em, Tag 'em, Graph 'em" (GTG) algorithm identified six cloud clusters within the West Africa region. However, the focus remains on the one that affected Burkina Faso. The graphical representation of that feature (Figure 5.8) indicates a relatively complex mesoscale convective system (MCS) feature throughout the duration of the time-period analyzed. The GTG algorithm recognizes the 'formation' of an MCS in the location 0000 UTC 31 Aug 2009 at the start of the period of the data analyzed. The shape and temperature criteria conform to that of a MCC at 1100 UTC, and the minimum duration of this area, and temperature criteria in order to be classified as a MCC, is met by 1700 UTC. At 1500 UTC 31 Aug 2009, the feature area was

251,744 km^2. The feature existed for 21 hours, until 0900 UTC 1 Sep 2009 (F34), meeting the area and temperature criteria. The noted initiated MCC and duration, according to the "Grab 'em, Tag 'em, Graph 'em," correlates with the times in Flooding in West Africa (2011) Module on COMET.

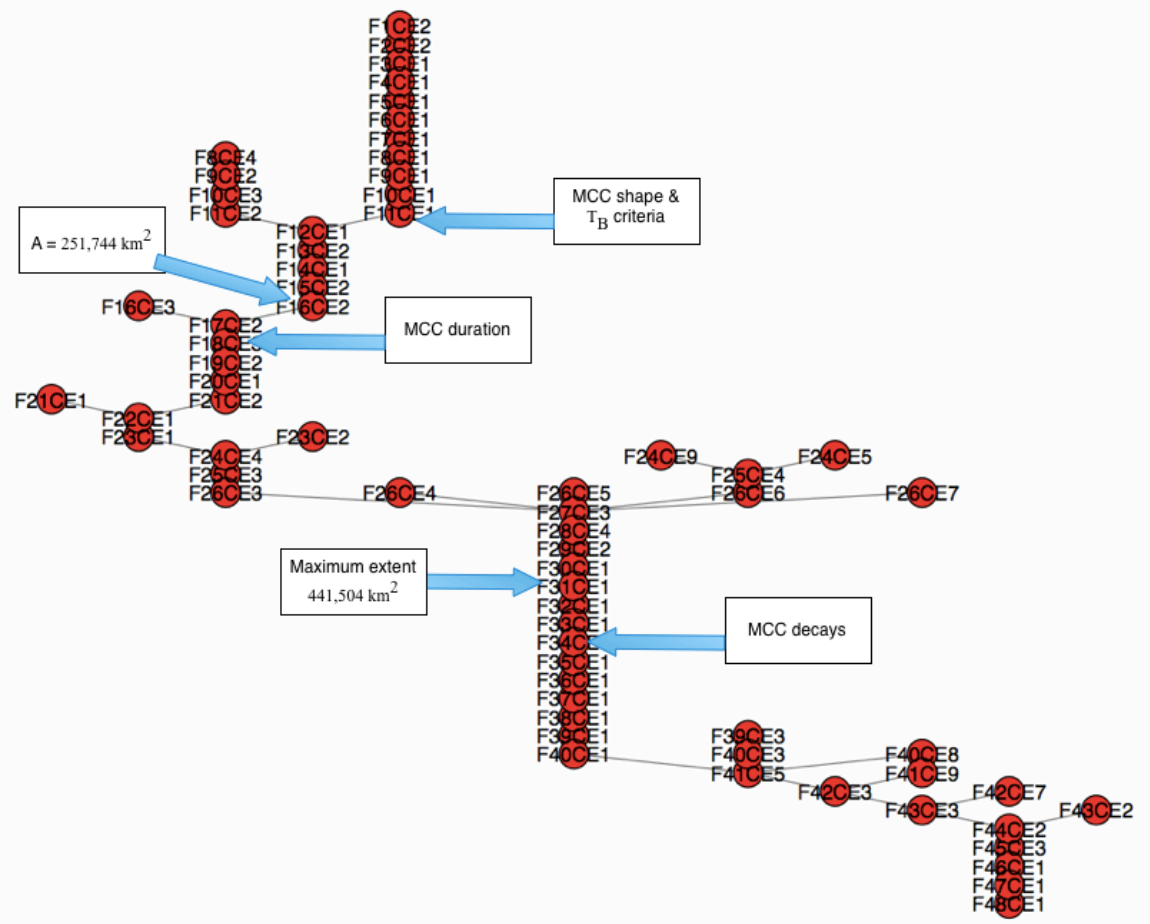


Figure 5.8. The graph representation of the mesoscale convective system observed from the "find cloud clusters" part of the GTG algorithm.

During this MCC's lifetime, the maximum extent was observed at 0600 UTC 1 Sep 2009 (F31CE1) as 441,504 km^2, and the eccentricity at this time was 0.54. The largest eccentricity recorded during the MCC feature was 0.6 at 1800 UTC 31 Aug 2009 and the feature area was 322,752 km^2. From the meteorological analysis, it is known that the feature was imbedded in a larger feature. The "Grab 'em, Tag 'em, Graph 'em" (GTG) indicates that the system 'decayed' between 1000 UTC 1 Sep 2009 (F35) and F42, and then strengthened, meeting the MCC area and temperature criteria, between 1800 UTC (F43) and 2100 UTC (F46). Figure 5.8 provides the full area distribution for the MCC.

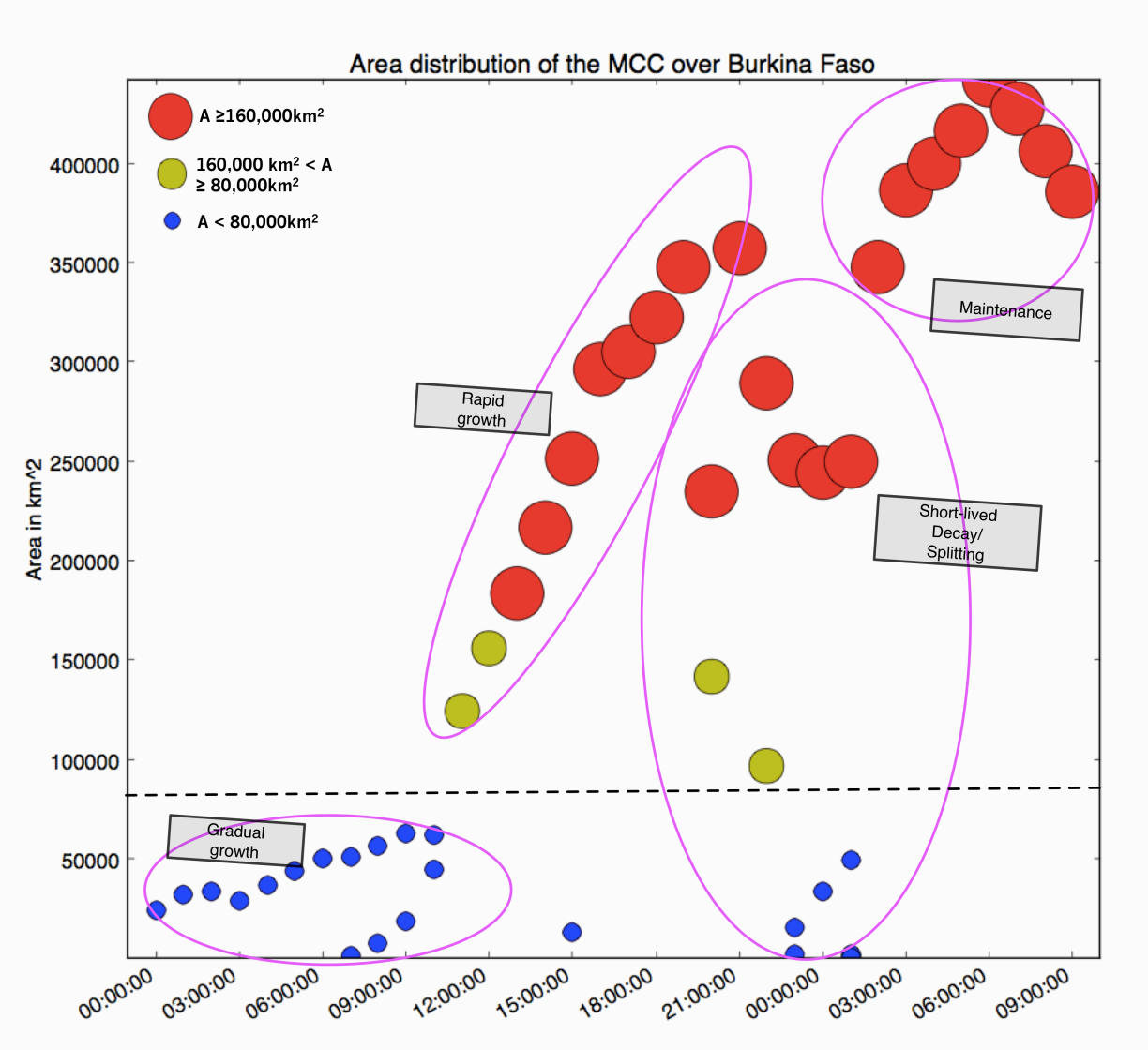


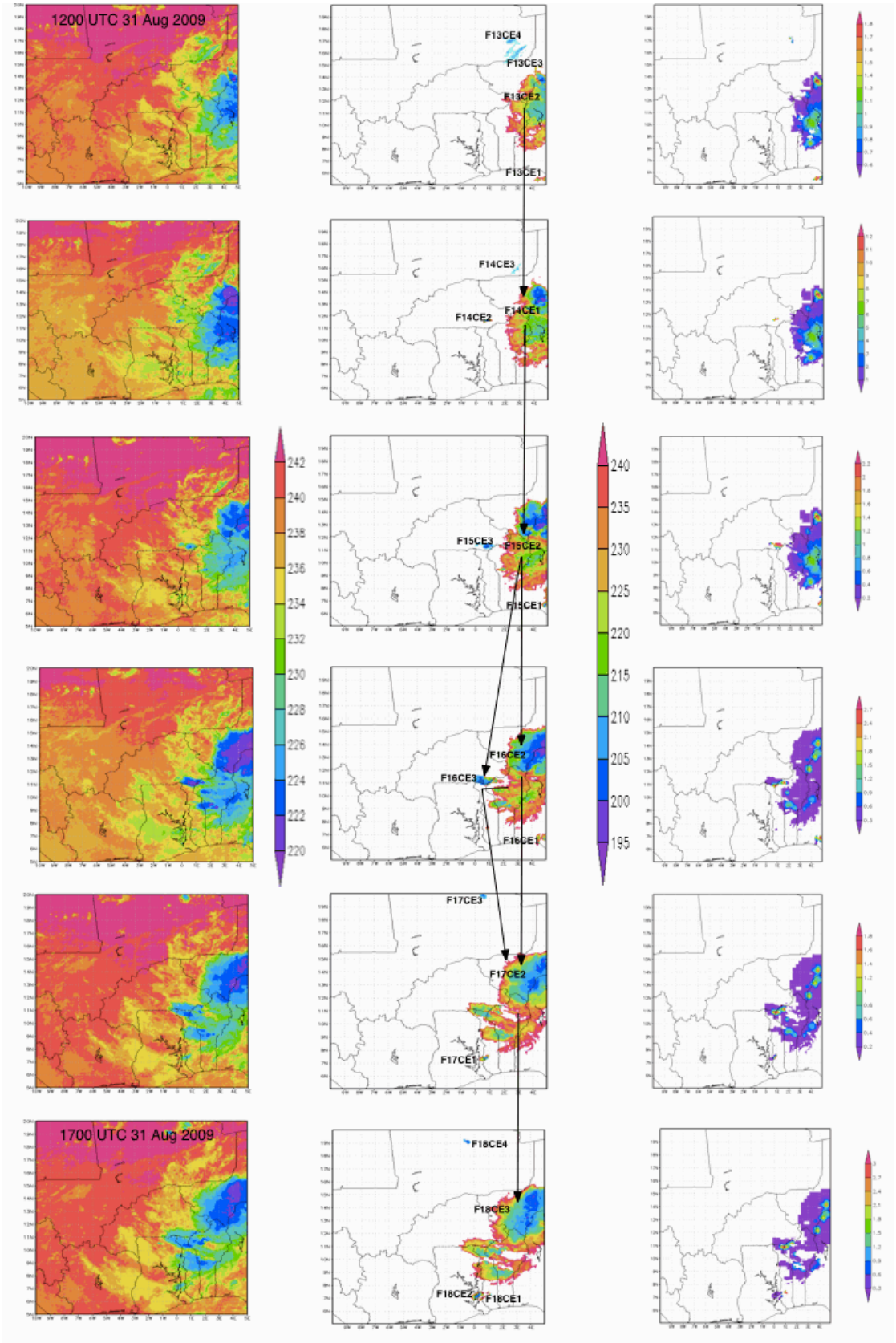
Figure 5.9. The area distribution for the MCC observed over Burkina Faso between 0000 UTC 31 Aug 2009 and 2300 UTC 1 Sep 2009. The dots qualitatively represent the area of each cloud element.

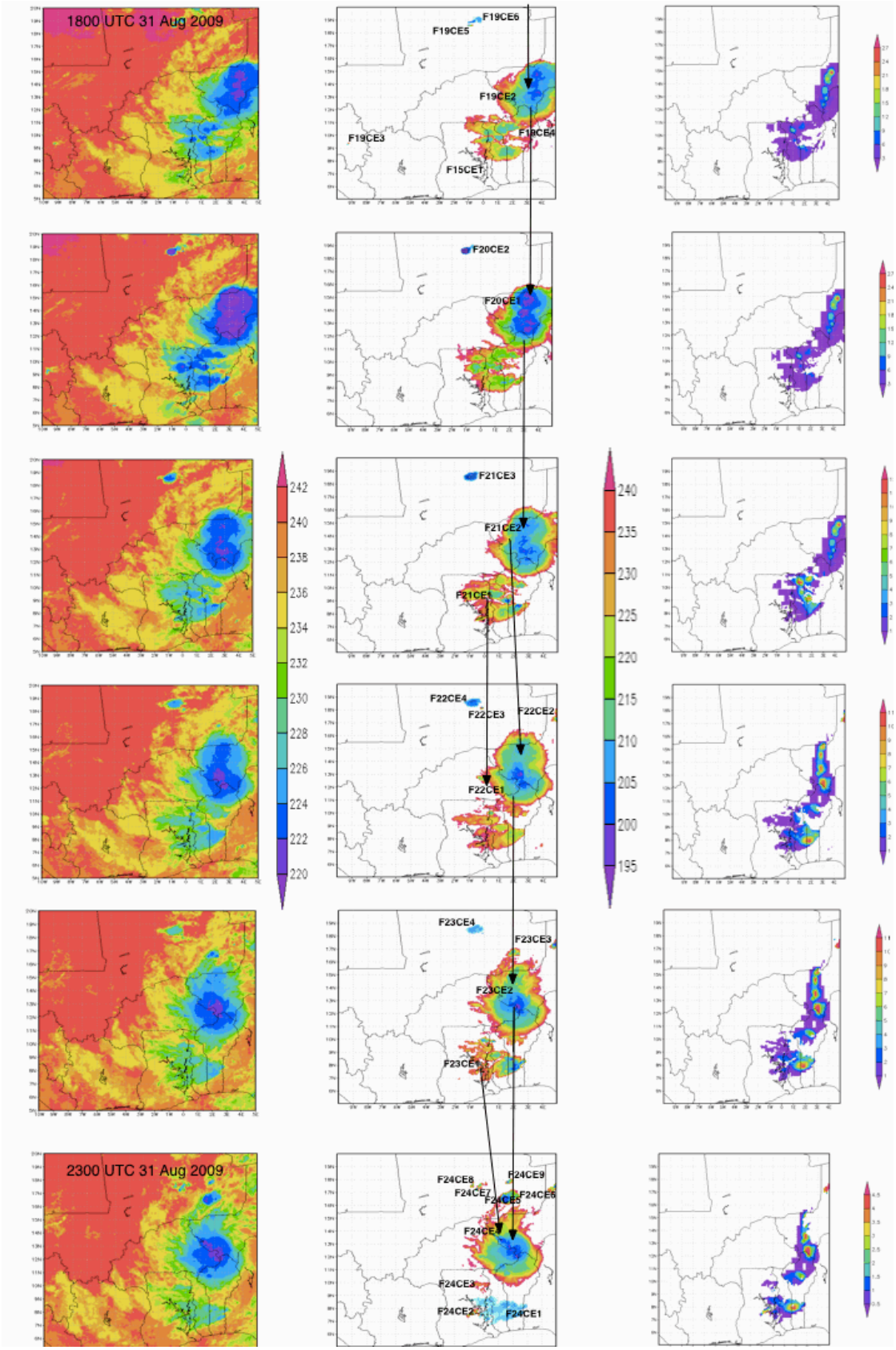
It is observed that the MCC feature gradually grew during until ~1200 UTC 31 Aug 2009, when it rapidly grew in size. There was a brief decay of the feature before regrowth, and maintenance of the large feature.

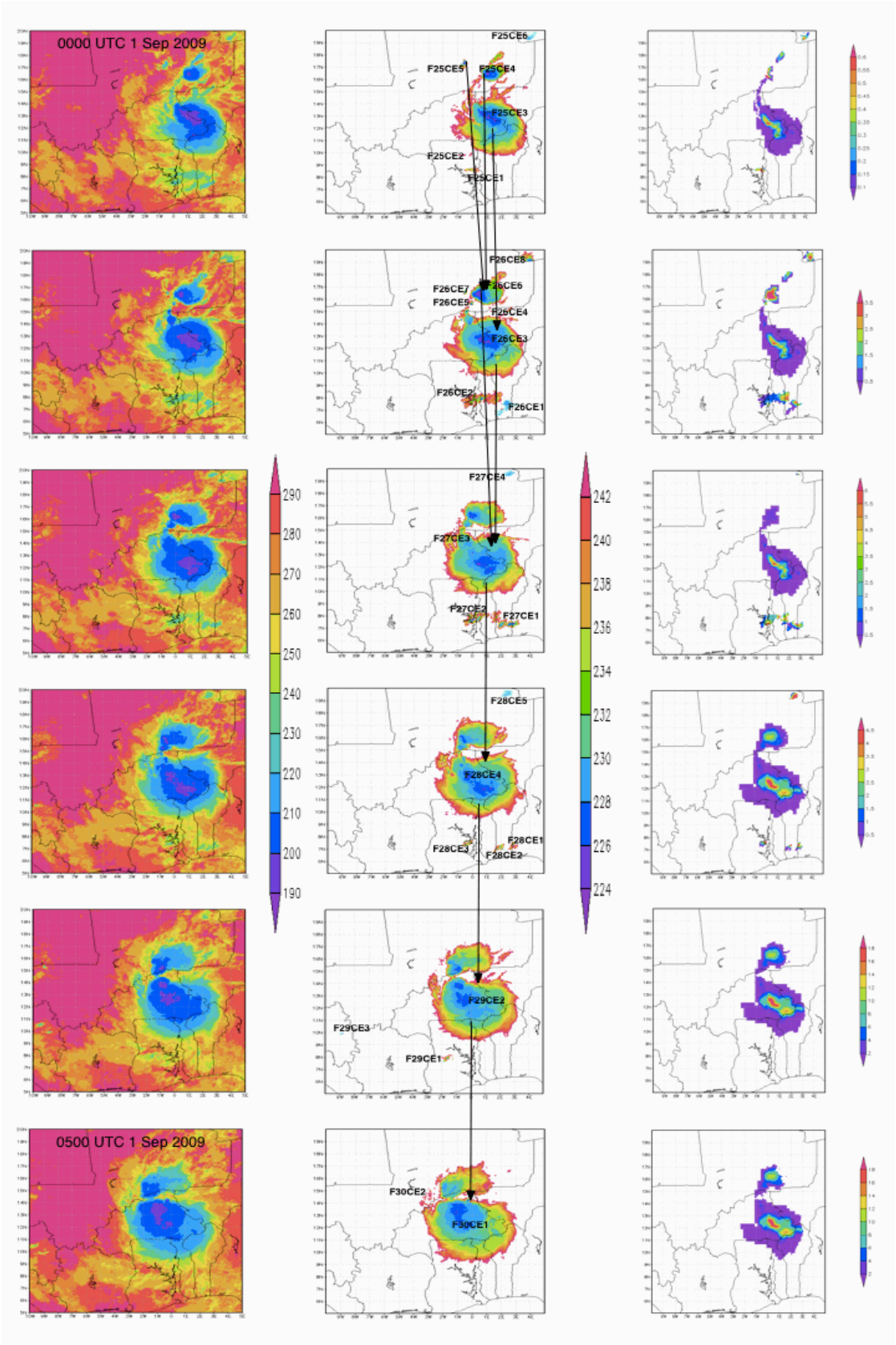
5.2.2 Precipitation Characteristics of the Mesoscale Convective Complex using the "Grab 'em, Tag 'em, Graph 'em" Method with Tropical Rainfall Measuring Mission Data

The original infrared images, the "Grab 'em, Tag 'em, Graph 'em" (GTG) algorithm cloud detection, and the Tropical Rainfall Measuring Mission (TRMM) rainfall are provided for select times within the mesoscale convective system in

Figure 5.10. Small areas are noticeable, being recognized as cloud elements in spite of their small sizes, as compared to the major feature. In general, the meteorological features as explained are represented within the data captured by the "Grab 'em, Tag 'em, Graph 'em" algorithm.


1200 UTC 31 Aug 2009
F13CE4
F13CE3
F13CE2
F13CE1
F14CE3
F14CE2
F14CE1
F15CE3
F15CE2
F15CE1
F16CE2
F16CE3
F16CE1
F17CE3
F17CE2
F17CE1
1700 UTC 31 Aug 2009
F18CE4
F18CE3
F18CE2
F18CE1
242
240
238
236
234
232
230
228
226
224
222
220
240
235
230
225
220
215
210
205
200
195


1800 UTC 31 Aug 2009
F19CE6
F19CE5
F19CE2
F19CE4
F19CE3
F15CE1
F20CE2
F20CE1
F21CE3
F21CE2
F21CE1
F22CE4
F22CE3
F22CE2
F22CE1
F23CE4
F23CE3
F23CE2
F23CE1
2300 UTC 31 Aug 2009
F24CE8
F24CE9
F24CE7
F24CE6
F24CE5
F24CE4
F24CE3
F24CE2
F24CE1
242
240
238
236
234
232
230
228
226
224
222
220
240
235
230
225
220
215
210
205
200
195


0000 UTC 1 Sep 2009
F25CE6
F25CE5
F25CE4
F25CE3
F25CE2
F25CE1
F26CE8
F26CE7
F26CE6
F26CE5
F26CE4
F26CE3
F26CE2
F26CE1
F27CE4
F27CE3
F27CE2
F27CE1
F28CE5
F28CE4
F28CE3
F28CE2
F28CE1
F29CE2
F29CE3
F29CE1
0500 UTC 1 Sep 2009
F30CE2
F30CE1
290
280
270
260
250
240
230
220
210
200
190
242
240
238
236
234
232
230
228
226
224

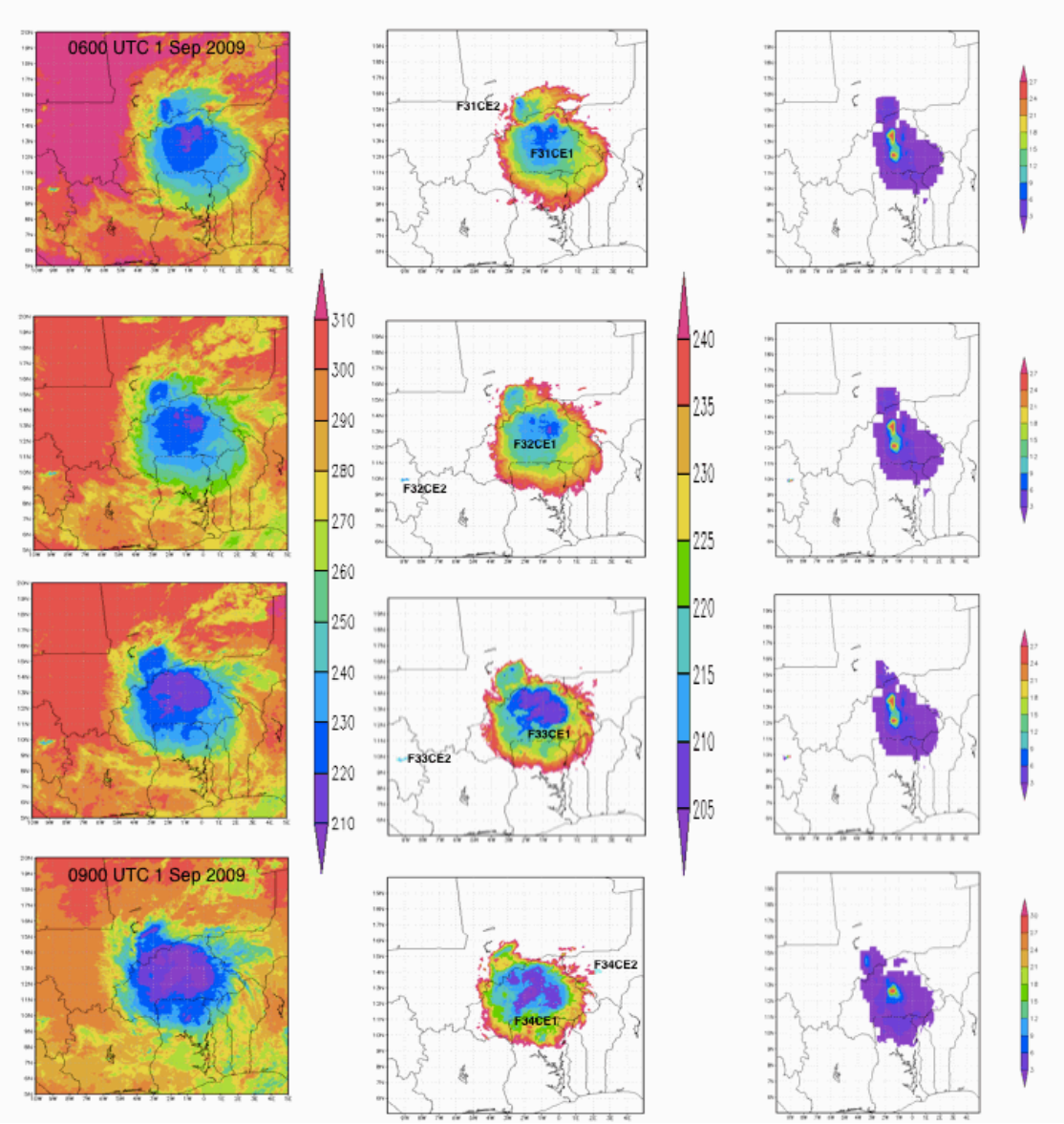


Figure 5.10. Cloud detection and tracking of the MCS centered over Burkina Faso from 1200 UTC 31 Aug 2009 to 0900 UTC 1 Sep 2009. The first column outlines the MERG images, the second column represents the MCSs detected by the GTG algorithm presented, and the third column illustrates the TRMM rainfall under the features as detected by the GTG algorithm. The black lines indicate some of the connectivity within the MCS.

The precipitation distribution (Figure 5.11) indicates that from early in the lifetime of the feature (the blue), a large percentage of the feature was precipitating. Furthermore, even in the moments of 'reorganization,' the feature was still precipitating over a large area.

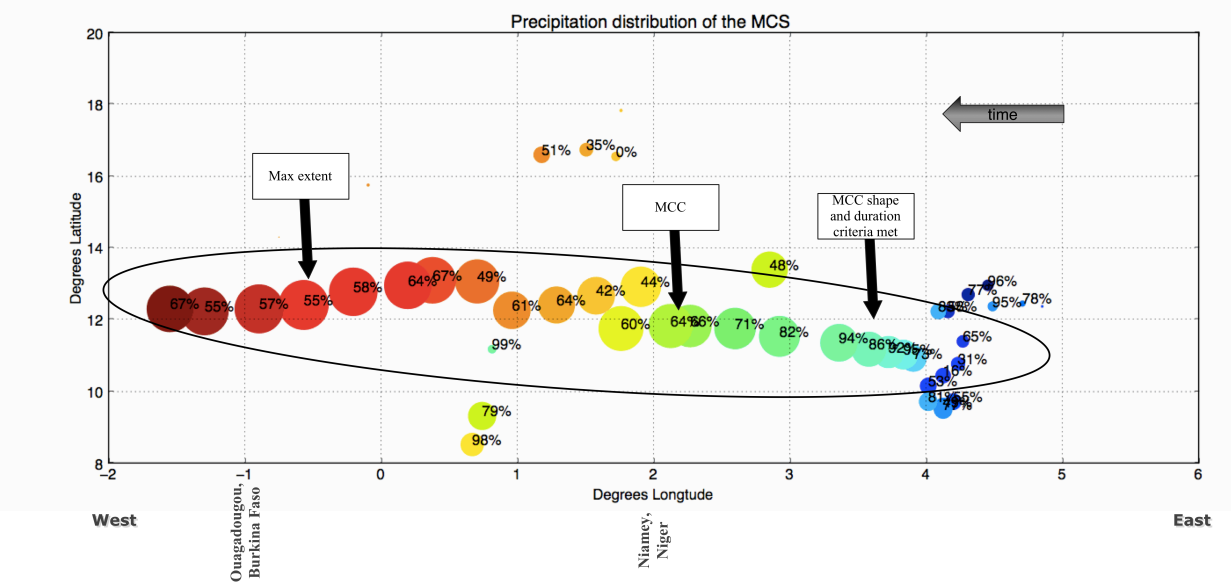
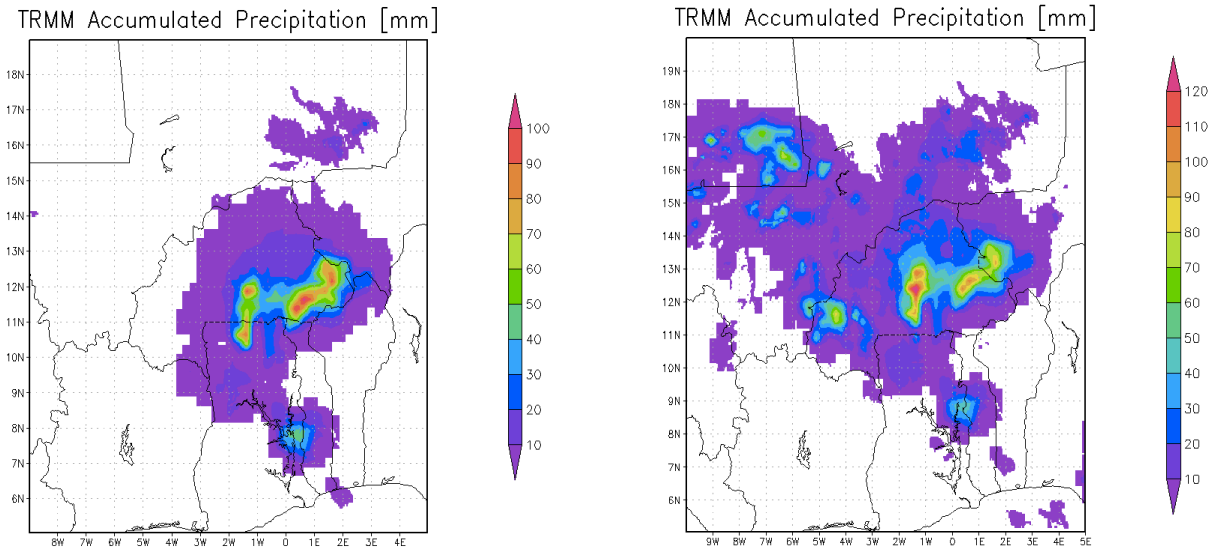


Figure 5.11. The spatial and temporal distribution of rainfall for cloud element in the MCS between 0000 UTC 31 Aug 2009 (blue circle) until 0900 UTC 1 Sep 2009 (red circle). The circles represent the relative area of each CE to the total area of the system. The percentage of the of the cloud element that was precipitating is illustrated. The oval indicates the main feature that grows into the MCC.

Figure 5.12 provides the accumulations for the feature between 0000 UTC to 0900 UTC 1 Sep 2009 and for the 24-hour period 0000 UTC 1 Sep 2009 to 0000 UTC 2 Sep 2009. As expected, the TRMM accumulations are lower than the rain gauge value reported. Nonetheless, it can be seen that most of the rainfall within the 24-hour period was associated with the MCC.



(a) (b)

Figure 5.12. (a) TRMM accumulations for the MCS feature between 0000 UTC and 0900 UTC 1 Sep 2009. (b) TRMM accumulations for the 24-hour period starting 0000 UTC 1 Sep 2009.

5.2.3 Summary of Case Study 2

The "Grab 'em, Tag 'em, Graph 'em" (GTG) algorithm was able to identify a mesoscale convective complex (MCC) and provide rainfall characteristics of the event. As this particular MCC was embedded within the large-scale synoptic feature of the African Easterly Wave, it demonstrates the ability of the GTG algorithm to identify MCC features even if they are embedded within other large-scaled mesoscale convective systems.

5.3 Case Study 3: A Long-term Analysis of Mesoscale Convective Complexes in West Africa

The objective of this case study is to determine the validity of the "Grab 'em, Tag 'em, Graph 'em" (GTG) algorithm being used in as an analysis tool to quantify relationships between cloud parameters and their precipitation characteristics for periods longer than the lifetime of a given feature. The experiment considers the domain 7.5°N–17.5°N and 20°E–17.5°W. The 4-km MERG dataset resolution implies that of 391,678 spatial data points are considered for each day and dataset - 1,041 points in the x-direction and 277 data points in the y-direction. This translates to 9,400,272 data points to be analyzed for each day. Whereas the short-term record experiments were conducted on a personal portable computer, these long-term experiments were conducted on a server system located at the National Aeronautics and Space Administration's (NASA) Jet Propulsion Laboratory (JPL). The MCC-Server is a dedicated virtual machine (VM) instance, supported by a fully redundant 8Gb/s Storage Area Network (SAN) system. This VM is hosted on the latest Dell enterprise-level hardware that includes redundant processors, redundant memory (mirrored), and direct access to the internal SAN switch fabric. The MCC-Server instance is deployed with 6 2.3Ghz CPU cores, 12GB 1600Mhz memory, and 2TB of SAN storage RAID disk space. Monthly runs took approximately 9 hours to complete.

The Laing *et al.* (1999) study determined from manual methods that 41 MCCs that occurred in West Africa between 17 July and 30 September 1987 – 10 in July, 19 in August and 12 in September – and characterized the precipitation associated with the features using microwave, infrared and surface gauge data. It should be noted that even though this is the only seasonal study of MCCs in West Africa, the year 1987 was a year of notable drought in the region (Lélé & Lamb 2010). The "Grab 'em, Tag 'em, Graph 'em" (GTG) algorithm identified 62 MCCs

between July - September 2006 (Table 5.1). This number of features identified is larger than the season according to the Laing *et al.* (1999) findings the July and August results are comparable, furthermore the total are comparable with data from Laing and Fritsch (1993). The maximum duration identified was 37 hours and the minimum duration 6 hours. (There was no limit to the length of time that the MCC criteria could be maintained in the experiment). The shape of the identified MCCs ranged between 0.7 and 1.0 for the identified features, with an average eccentricity of 0.85. The average area of the features is within the limits of the MCC size limits for this location (Laing & Fritsch, 1993). The rainfall rates are between the limits of the aforementioned studies, though it is recognized the actual values may be higher.

Table 5.1. The cloud and rainfall variables analyzed in the "Grab 'em, Tag 'em, Graph 'em" algorithm for Jul – Sep 2006.

Symbols	***Definition and units***	***Jul***	***Aug***	***Sep***
N	Number of features	23	20	19
LD	Life duration (h)	10.04	11.0	12.35
V_{av}	Propagation speed (m s^{-1})	9.06	9.25	8.38
$A_{IR_{max}}$	Average maximum area (km^2) from IR data	225,644.67	285, 687.20	261,709.16
$A_{TRMM_{max}}$	Average precipitating area (km^2) from TRMM data	121,913.89	146,637.09	156,569.55
P_{max}	Average maximum precipitation rate (mm h^{-1})	22.85	23.64	20.64
pA	Average area precipitating w.r.t. the area (%)	62.09	68.57	75.35

5.3.1 Summary of Case Study 3

The results from this experiment encourage a more thorough examination of the algorithm for periods and domains longer and larger than individual events. But the "Grab 'em, Tag 'em, Graph 'em" method has not yet been optimized for such studies. There were some computational shortcomings that were identified during the case study as will be explained in the following section.

5.4 Computational Lessons

The short-term datasets highlighted the need to generate better storage methods such as data type compression methods for generated files. As an illustration, the MERG raw data files represent the brightness temperature as an integer, by subtracting '75' from the original floating-point number. This type of compression may be useful to consider for future development.

During the case study analyzing long-term dataset, limitations regarding the GTG algorithm's handling of computer resources were experienced (Figure 5.13). Specifically, the MCC-Server consistently terminated the program due to a lack of memory. These crashed lead to experimenting with the domain and amount of data being ingested to identify the core reasons for the crashes. The exercises identified two causes, though these may not be the only causes. The first reason proposed is that a large percentage of computational resources are spent managing data files on the front end in order to build the graph. This limitation was demonstrated in the larger domain runs, for example the domain 7.5°N–17.5°N and 20°E–17.5°W where the GTG algorithm terminated after opening 47 days of data.

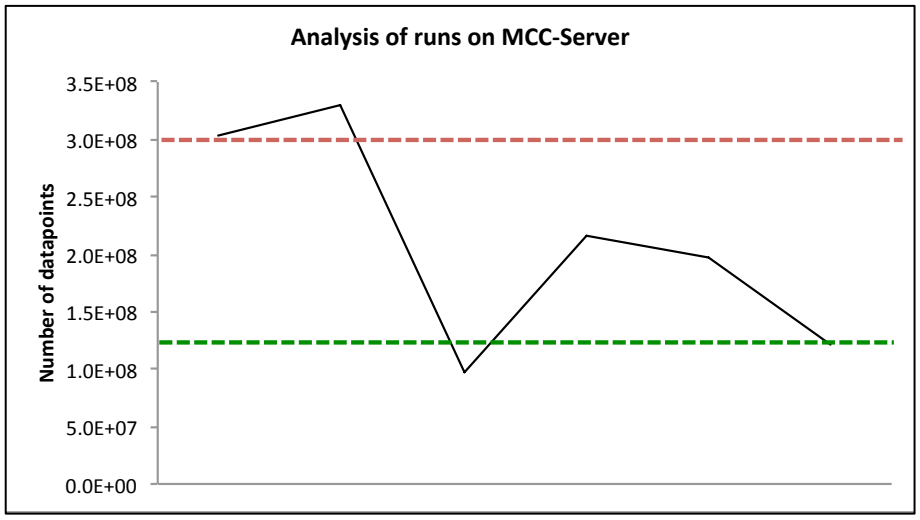


Figure 5.13. An analysis of the limitations of the number of data points that can be accessed using the "Grab 'em, Tag 'em, Graph 'em" algorithm for long-term runs. The red line indicates the data point limit due to data extraction, while the green line indicates the data point limit related to data analysis.

The second limitation was identified memory leaks during the "finding cloud clusters" and /or "finding cloud elements" stage. In this case, it is recognized that GTG algorithm has not been tested for optimum datatype implementation and data compression methods. To illustrate, using a 64-bit floating-point number to hold a variable that could be represented as a 16-bit integer, inefficiently uses the computer resources. Furthermore, it is noted that the GTG algorithm code is not yet parallelized to leverage the multiprocessor capabilities of systems.

CHAPTER 6. CONCLUSIONS, CONTRIBUTIONS AND FUTURE WORK

Time- and computer-efficient methods for identifying weather systems in long-term high spatial and temporal resolution datasets are required in the Atmospheric Science community. The February 2014 launch of the Global Precipitation Measurement (GPM) satellite will improve on the Tropical Rainfall Measuring Mission (TRMM) product and provide greater coverage between 65 °N and 65 °S. The GPM satellite will also generate high-resolution data in both space and time. Even before this dataset amasses, methods for efficiently searching for weather features from satellite datasets will be necessary. The "Grab 'em, Tag 'em, Graph 'em" (GTG) algorithm begins to address the need for such methods in gridded four dimensional (time, lat, lon, variable) datasets.

6.1 Conclusions

The first step to addressing this need was presented in the work conducted in this dissertation. This research considered a method for identifying mesoscale convective complexes (MCCs) in infrared and precipitation satellite datasets of varying resolutions, through the implementation of methods uncommon to those existing in the atmospheric science data processes. The method, the "Grab 'em, Tag 'em, Graph 'em" (GTG) algorithm, was implemented using graph theory, and Python as the programming language. The key findings of the GTG algorithm were:

1. The GTG algorithm can identify mesoscale convective systems (MCSs) in infrared datasets, as well as efficiently and accurately track them through time.
2. The GTG algorithm can identify mesoscale convective complexes (MCCs) – a sub-class of large-scale MCSs.

3. The GTG algorithm can complete an end-to-end identification of MCC features in multiple datasets without manual intervention.
4. The GTG algorithm can easily be deployed as a standalone program on a personal computing device for short-term data analysis.
5. The GTG algorithm can be easily adapted to identify other large-scale MCSs, allow for other metrics calculations, and provide visualizations.
6. The GTG algorithm has been a developed in a modular fashion thus allowing for code reusability and encourages development.

6.2 Contributions

6.2.1 Thesis contributions

Overall, the work completed in this dissertation has provided a set of methods, algorithms and techniques for searching four-dimensional datasets. The key contributions of this work extend beyond the Atmospheric Science community, and extend into the Computer Science community as well. From the studies presented, the key contributions include:

1. A novel method of considering the identifying weather features. The "Grab 'em, Tag 'em, Graph 'em" (GTG) expounds on the premise of 'grabbing' only the areas of interest required from a data frame for analysis and graphing them according to their spatial correlations between times. This directly contrasts the static-in-space domain approaches that are commonly used in the Atmospheric Sciences, to analyze weather features and / or climate features.
2. An automated method for identifying weather features in infrared satellite datasets while following their complex evolution for timescales longer than an individual event.

3. A fully automated process of finding the associated precipitation of identified MCC (and large-scaled MCSs) in various datasets to provide precipitation characteristics.
4. A search and clip functionality that should work with any four dimensional (time, lat, lon, variable) dataset.
5. A modification of traditional graph traversal methods to address a particular application. The modified depth-first iterative deepening traversal is used to traverse the tree to determine any mesoscale convective complexes (MCCs).

6.2.2 Research Accomplishments

In addition to the contributions listed above, the following practical accomplishments related to this work have been completed to date:

1. Establishing of collaborations between the Regional Climate Model Evaluation System (RCMES) team and Howard University's Atmospheric Science Program.
2. Publishing of papers including Whitehall *et al.*, (2012), Whitehall, Chiao & Mayers-Als, (2013), and Mattmann *et al.*, (2013)
3. Presentations of posters at the American Geophysical Union (AGU) Fall Meeting in 2012 and 2013, as well as an oral presentation at the American Meteorological Society (AMS) 94th Annual Meeting in February 2014. The poster presentation titled "An Automated Method to Identify Mesoscale Convective Complexes (MCCs) Implementing Graph Theory" at the AGU Fall 2013 meeting won an Earth and Space Science Informatics Outstanding Student Presentation Award.
4. Publishing of aspects of the "Grab 'em, Tag 'em, Graph 'em" algorithm on the Apache Open Climate Workbench project wiki.

5. Publishing the "Grab 'em, Tag 'em, Graph 'em" source code as a branch on the Apache Open Climate Workbench GitHub repository (https://github.com/kwhitehall/climate)

6.3 Practical Concerns

A concern about the efficiency of the "Grab 'em, Tag 'em, Graph 'em" (GTG) algorithm was outlined, where the algorithm labored with analyses of more than 50 days of data at a time. While this seems to potentially reduce the applicability of the GTG algorithm with long-term data analysis, there are some immediate future research points that are geared at addressing this concern.

6.4 Future Work

The findings of this work have shown promise of the "Grab 'em, Tag 'em, Graph 'em" (GTG) algorithm in applications requiring short-term or long-term satellite datasets. As such, there are many fruitful areas of research that are identified.

6.4.1 Immediate Future Work

In the immediate future, addressing the practical concerns / limitations observed in this study is necessary. As such, suggested near-term research directions include:

1. Further testing of single-event and / or short-term case studies in West Africa and other locations for full verification of the algorithm in infrared satellite datasets would further the algorithm's usefulness in the community.
2. Exploring the precipitation characteristics of MCCs in various datasets e.g. GPCC, CMORPH and reanalysis datasets.

3. Analysis of the identified systems from the summer 2006 case study for full verification of the long-term implementation.
4. Extensive debugging to improve the use of computer resources, through the optimization of datatypes and objects used in the GTG algorithm, as well as the implementation of datatype compression for storage, following the format of the original datasets.
5. Extending the GTG algorithm to searches within databases of satellite data. Furthermore, using the Regional Climate Model Evaluation Database (RCMED) as an instance of a "non-graph" database, and creating a Neo4j[1] database for an instance of a graph database, encourages the open-sourced development of the algorithm.
6. Exploring the algorithm design and computational efficiency of the modified depth-first iterative deepening traversal.
7. Exploring the range of mesoscale convective systems the GTG algorithm can capture.

6.4.2 Future Directions

The potential of the "Grab 'em, Tag 'em, Graph 'em" (GTG) algorithm in searching satellite datasets can be applied to both weather and climate applications. One application for the GTG algorithm includes using it to identify mesoscale convective complexes (MCCs) in weather model datasets, such as those from the Weather, Research and Forecasting (WRF) model. Another application involves using the GTG algorithm to characterize precipitation characteristics of MCCs from various satellite instruments. Additionally,

1. Parallelizing the GTG algorithm, thus increasing the code run-time efficiency and usage of computation resources. This will enhance the use of the GTG algorithm with

[1] http://www.neo4j.org/

long-term and / or high-resolution datasets. Parallelization may include splitting the tasks on computer processors according to times within the dataset being analyzed, or according to each cloud cluster formation identified. Parallelization may also include finding ways to scale the graph through investigating the use of graph processing engines like the Apache Hadoop[2]-based Apache Giraph[3].

2. Explore using the GTG algorithm with weather model outputs for forecasting applications.
3. Explore using the GTG algorithm in climate evaluations.
4. Exploring methods to visualize the results from long-term records of data
5. Exploring using machine learning with the algorithm to forecast MCCs.

As the future work listed above demonstrates, this thesis enables the beginning of new research areas, especially related to identifying and characterizing weather features in big datasets.

[2] http://hadoop.apache.org/
[3] https://giraph.apache.org/

BIBLIOGRAPHY

Adler, R. F., & Negri, A. J. (1988). A satellite infrared technique to estimate tropical convective and stratiform rainfall. *Journal of Applied Meteorology*, *27*(1), 30-51.

Adler, R. F., Negri, A. J., Keehn, P. R., & Hakkarinen, I. M. (1993). Estimation of monthly rainfall over Japan and surrounding waters from a combination of low-orbit microwave and geosynchronous IR data. *Journal of Applied Meteorology*, *32*(2), 335-356.

Aggarwal, C. C., & Wang, H. (2010). *Managing and mining graph data* (Vol. 40). Springer.

Anderson, C. J. & Arritt, R. W. (1998). Mesoscale convective complexes and persistent elongated convective systems over the United States during 1992 and 1993. *Monthly Weather Review, 126*(3), 578-599.

Arkin, P. A. (1979). The Relationship between Fractional Coverage of High Cloud and Rainfall Accumulations during GATE over the B-Scale Array. *Monthly Weather Review, 107*, 1382-1387.

Arnaud, Y., Desbois, M., & Maizi, J. (1992). Automatic Tracking and Characterization of African Convective Systems on Meteosat Pictures. *Journal of Applied Meteorology, 31*(5), 443-453.

Augustine, J. A., & Howard, K. W. (1991). Mesoscale convective complexes over the United States during 1986 and 1987. *Monthly Weather Review, 119*, 1575-1589.

Beer, T., Greenhut, G. K., & Tandoh, S. E. (1977). Relations between the Z Criterion for the Subtropical High, Hadley Cell Parameters and Rainfall in Northern Ghana. *Monthly Weather Revie*w, *105*(7), 849-855.

Blamey, R. C., & Reason, C. J. (2012). Mesoscale Convective Complexes over Southern Africa. *Journal of Climate, 25*, 753-766.

Bouniol, D., Delanoë, J., Duroure, C., Protat, A., Giraud, V., & Penide, G. (2010). Microphysical characterisation of West African MCS anvils. *Quarterly Journal of the Royal Meteorological Society, 136*(S1), 323-344.

Burkina Faso: 2009 Flooding Situation Report No. 1. (2009, September 4). *United Nations Office for the Coordination of Humanitarian Affairs.* Retrieved from http://reliefweb.int/report/burkina-faso/burkina-faso-2009-flooding-situation-report-no1.

Burpee, R. W. (1972). The origin and structure of easterly waves in the lower troposphere of North Africa. *Journal of Atmopheric Science, 29*, 77-90.

Burpee, R. W. (1974). Characteristics of North African Easterly Waves During the Summers of 1968 and 1969. *Journal of Atmospheric Sciences, 31*, 1556-1570.

Capriolo, E., Wampler, D., & Rutherglen, J. (2012). *Programming Hive*. O'Reilly

Carvalho, L. M., & Jones, C. (2001). A Satellite Method to Identify Structural Properties of Mesoscale Convective Systems Based on the Maximum Spatial Correlation Tracking Technique (MASCOTTE). *Journal of Applied Meteorology, 40*, 1683-1701.

Charney, J. G. (1975). Dynamics of deserts and drought in the Sahel. *Quarterly Journal of the Royal Meteorological Society, 101*(428), 193-202.

Chaudhuri, S., & Middey, A. (2009). Applicability of bipartite graph model for thunderstorms forecast over Kolkata. *Advances in Meteorology*, 1–12.

Chaudhuri, S., & Middey, A. (2011). Nowcasting thunderstorms with graph spectral distance and entropy estimation. *Meteorological Applications, 18*(2), 238–249.

Cook, K. H. (1999). Generation of the African Easterly Jet and Its Role in Determining West African Precipitation. *Journal of Climate, 12*, 1165-1184.

Crichton, D., Mattmann, C., Cinquini, L., Braverman, A., Waliser, D., Hart, A., et al. (2012). Sharing Satellite Observations with the Climate Modeling Community: Software and Architecture. *Software, IEEE, 29*(5), 63-71.

D'Amato, N., & Lebel, T. (1998). On The Characteristics of the Rainfall Events in the Sahel with a View to the Analysis of Climatic Variability. *International Journal of Climatology, 18*, 955-974.

Desbois, M., Kayiranga, T., Gnamien, B., Guessous, S., & Picon, L. (1988). Characteristics of Some Elements of the Sahelian Climate and Their Interannual Variations for July 1983, 1984 and 1985 from the Analysis of METEOSAT ISCCP Data. *Journal of Climate, 1*, 867-904.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*,*1*(1), 269-271.

Dixon, M., & Wiener, G. (1993). TITAN: Thunderstorm Identification, Tracking, Analysis and Nowcasting – A Radar-based Methodology. *Journal of Atmospheric and Oceanic Technology, 10*(6), 785–797.

Duvel, J. P. (1989). Convection over Tropical Africa and the Atlantic Ocean during Northern Summer. Part I: Interannual and Diurnal Variations. *Monthly Weather Review, 117*, 2782-2799.

El Houssein, T. A., & Decleir, H. (1998). Upper air characteristics of wet and dry days in tropical West Africa during the rainy season. *Water Resources Variability in Africa during the XXth Century*, *252*, 55-62.

Eltahir, E. A., & Gong, C. (1996). Dynamics of wet and dry years in West Africa. *Journal of Climate, 9*, 1030-1042.

Fiolleau, T., & Roca, R. (2013). An algorithm for the detection and tracking of tropical mesoscale convective systems using infrared images from geostationary satellite. *IEEE Transactions on Geoscience and Remote Sensing, 51*(7), 1-4.

Flooding in West Africa. (2011). *University Corporation for Atmospheric Research.* Available from http://www.meted.ucar.edu/asmet/w_africa/print.htm.

Floods displace thousands in Burkina Faso. (2009, September 10). *Tearfund.* Retrieved from http://reliefweb.int/report/burkina-faso/floods-displace-thousands-burkina-faso.

Fontaine, B., & Janicot, S. (1992). Wind field coherence and its variations over West Africa. *Journal of Climatology, 5*, 512-524.

Fontaine, B., & Janicot, S. (1996). Sea Surface Temperature Fields Associated with West African Rainfall Anomaly Types. *Journal of Climate, 9*(11), 2935-2940.

Fontaine, B., Garcia-Serrano, J., Roucou, P., Rodriguez-Fonseca, B., Losada, T., Chauvin, F., et al. (2010). Impact of warm and cold situations in the Mediterranean basins on the West African monsoon: observed connection patterns (1979 - 2006) and climate simulations. *Climate Dynamics, 35*, 95-114.

Frequently Asked Questions. (n.d.). *World Meteorological Organization.* Retrieved from http://www.wmo.int/pages/prog/wcp/ccl/faqs.html.

Fritsch, J. M., Kane, R. J., & Chelius, C. H. (1986). The contribution of mesoscale convective weather systems to the warm season precipitation in the United States. *Journal of Climate and Applied Meteorology, 25*, 1333-1345.

Galvin, J. F. (2010). Two easterly waves in West Africa in summer 2009. *Weather, 65*(8), 219–227.

Gamache, J. F., & Houze Jr., R. A. (1983). Water Budget of a Mesoscale Convective System in the Tropics. *Journal of Atmospheric Sciences, 40*, 1835-1850.

García-Herrera, R., Barriopedro, D., Hernández, E., Paredes, D., Correoso, J., & Prieto, L. (2005). The 2001 mesoscale convective systems over Iberia and the Balearic Islands. *Meteorology and Atmospheric Physics, 90*, 225-243.

Giannini, A., Saravanan, R., & Chang, P. (2003). Oceanic Forcing of Sahel Rainfall on Interannual to Interdecadal Time Scales. *Science, 302*, 1027-1030.

Giorgi, F., Jones, C., & Asrar, G. R. (2009). Addressing climate information needs at the regional levels: the CORDEX framework. *WMO Bulletin , 58*(3), 175-183.

Goyens, C., Lauwaet, D., Schröder, M., Demuzere, M., & Van Lipzig, N. P. (2011). Tracking mesoscale convective systems in the Sahel: relation between cloud parameters and precipitation. *International Journal of Climatology, 32*, 1921-1934.

Gray, W. M., & Jacobson Jr., R. W. (1977). Diurnal variation of deep cumulus convection. *Monthly Weather Review, 105*, 1171-1188.

Hagberg, A. A., Schult, D. A., & Swart, P., J. (2008). Exploring network structure, dynamics, and function using NetworkX, (*Proceedings of the 7th Python in Science Conference (SciPy2008)*, Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), Pasadena, CA USA, 11–15, Aug 2008

Hamilton, R. A., & Archbold, J. W. (1945). Meteorology of Nigeria and adjacent territory. *Quarterly Journal of the Royal Meteorological Society, 71*, 231-265.

Han, J., Kamber, M., & Pei, J. (2006). *Data mining: concepts and techniques*. Morgan Kaufmann.

Hart, A., Goodale, C., Mattmann, C., Zimdars, P., Crichton, D., Lean, P., et al. (Eds.). (2011). A Cloud-Enabled Regional Climate Model Evaluation System. *ICSE 2011 Workshop on Software Engineering for Cloud Computing – SECLOUD.* Honolulu, HI.

Hartmann, D. L. (1994). *Global Physical Climatology* (Vol. 56). Waltham, MA: Academic Press.

Hastenrath, S. (1991). *Climate dynamics of the tropics.* Berlin, Germany: Kluwer Academic Publishers.

Hewitt, C. D., & Griggs, D. J. (2004). Ensembles-based Predictions of Climate Change and their Impacts. *Eos, 85*, 566.

Holton, J. A. (2004). *An Introduction to Dynamic Meteorology* (4th ed.). Waltham, MA: Elsevier.

Hoskins, B. (2013). The potential for skill across the range of the seamless weather-climate \ prediction problem: a stimulus for our science. *Quarterly Journal of the Royal Meteorological Society*, *139*(672), 573-584.

Houze Jr., R. A. (1993). *Cloud Dynamics.* Waltham, MA: Academic Press.

Houze Jr., R. A., Smull, B. F., & Dodge, P. (1990). Mesoscale organization of springtime rainstorms in Oklahoma. *Monthly Weather Review, 118*(3), 613-654.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 0090-95.

IPCC, 2007: Climate Change 2007: The Physical Science Basis. Contribution of Working Group

I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M.Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM*, *52*(8), 36-44.

Jenkins, G. S. (2000). TRMM satellite estimates of convective processes in Central Africa during September, October, November 1998: implications for elevated Atlantic tropospheric ozone. *Geophysical Research Letters, 27*(12), 1711-1714.

Jobard, I., & Desbois, M. (1992). Remote sensing of rainfall over topical Africa using Meteosat infrared imagery: sensitivity to time and space averaging. *International Journal of Remote Sensing, 13*(14), 2683-2700.

Jobard, I., & Desbois, M. (1994). Satellite estimation of the tropical precipitation using the METEOSAT and SMM/I data. *Atmospheric Research, 34*, 285-298.

Johnson, J. T., Mackeen, P. L., Witt, A., Mitchell, E. D., Stumpf, G. J., Eilts, M. D., et al. (1998). The storm cell identifcation and tracking alogrithm: An enhanced WSR-88D algorithm. *Weather Forecas , 13*(2), 263-276.

Korf, R. E. (1985). Depth-first iterative-deepening: An optimal admissible tree search. *Artifical Intelligence, 27*(1), 97-109.

Koteswaram, P. (1958). The Easterly Jet Stream in the Tropics. *Tellus, 10*(1), 43-57.

Kuettner, J. P. (1974). General description and central program of GATE. *Bulletin American Meteorological Society*, *55*(7), 712-719.

Laing, A.G., & Evans, J. L. (2011). *Introduction to Tropical Meteorology* (2nd ed.). Retrieved from http://www.goes-r.gov/users/comet/tropical/textbook_2nd_edition/index.htm.

Laing, A. G., & Fritsch, J. M. (1993). Mesoscale Convective Complexes in Africa. *Monthly Weather Review, 121*, 2254-2263.

Laing, A. G., & Fritsch, J. M. (1997). The global population of mesoscale convective complexes. *Quarterly Journal of the Royal Meteorological Society, 123*, 389-405.

Laing, A. G., & Fritsch, J. M. (2000). The large-scale environments of the global populations of mesoscale convective complexes. *Monthly Weather Review*, *128*(8).

Laing, A. G., Fritsch, J. M., & Negri, A. J. (1999). Contribution of mesoscale convective complexes to rainfall in Sahelian Africa: Estimates from geostationary infrared and passive microwave data. *Journal of Applied Meteorology*, *38*(7), 957-964.

Lamb, P. J. (1978). Case Studies of Tropical Atlantic Surface Circulation Patterns During Recent Sub-Saharan Weather Anomalies: 1967 and 1968. *Monthly Weather Review*, *106*, 482–491. http://dx.doi.org/10.1175/1520-0493(1978)106<0482:CSOTAS>2.0.CO;2

Laurent, H., D'Amato, N., & Lebel, T. (1998). How important is the contribution of the mesoscale convective complexes to the Sahelian rainfall. *Physics and Chemistry of the Earth , 23*, 629-633.

Le Barbé, L., Lebel, T., & Tapsoba, D. (2002). Rainfall Variability in West Africa during the Years 1950 - 90. *Journal of Climate, 15*, 187-202.

Leary, C. A., & Houze Jr., R. A. (1980). The Contribution of Mesoscale Motions to the Mass and Heat Fluxes of an Intense Tropical Convective System. *Journal of Atmospheric Sciences, 37*, 784-796.

Lebel, T., Taupin, J. D., & D'Amato, N. (1997). Rainfall monitoring during HAPEX-Sahel. 1. General rainfall conditions and climatology. *Journal of Hydrology*, *188*, 74-96.

Lélé, M., & Lamb, P.J. (2010). Variability of the Intertropical Front (ITF) and rainfall over the west African Sudan-Sahel Zone. *Journal of Climate, 23(14)*, 3984 – 4004.

Lindzen, R. S., & Nigam, S. (1987). On the role of Surface Tempeature Gradients in Forcing Low-level Winds and Convergence in the Tropics. *Journal of Atmospheric Science, 44*(17), 2418-2436.

Liou, K. N. (2002). *An introduction to atmospheric radiation* (2nd ed.). (Vol. 84). Waltham: Academic Press.

Machado, L. A., & Laurent, H. (2004). The convective system area expansion over Amazonia and its relationships with convective system life duration and high-level wind divergence. *Monthly Weather Review, 132*(3), 714-725.

Machado, L. A., Rossow, W. B., Guedes, R. L., & Walker, A. W. (1998). Life cycle variations of mesoscale convective systems over the Americas. *Monthly Weather Review, 126*, 1630-1654.

Maddox, R. A. (1980a). An Objective Technique for Separating Macroscale and Mesoscale Features in Meteorological Data. *Monthly Weather Review, 108*, 1108-1121.

Maddox, R. A. (1980b). Mesoscale convective complexes. *Bulletin of the American Meteorological Society, 61*, 1374-1387.

Mapes, B. E., & Houze Jr., R. A. (1993). Cloud clusters and superclusters over the oceanic warm pool. *Monthly Weather Review, 121*(5), 1398-1415.

Mathon, V., & Laurent, H. (2001). Life cycle of Sahelian mesoscale convective cloud sytems.

Quarterly Journal of the Royal Meteorological Society, 127, 377-406.

Mathon, V., Laurent, H., & Lebel, T. (2002). Mesoscale Convective System Rainfall in the Sahel. *Journal of Applied Meteorology, 41*, 1081-1092.

Mattmann, C. A. (2013). A vision for data science. *Nature, 493*, 473-475.

Mattmann, C. A., Crichton, D. J., Braverman, A., Williams, D., Gunson, M., Woollard, D., et al. (2009). A Distributed Computing Infrastructure for the Evaluation of Climate Models Using NASA Observational Data. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on* (pp. 231-232). IEEE.

Mattmann, C. A., Crichton, D., Hart, A., Kelly, S., Goodale, C., Downs, R. R., et al. (2012). Understanding Open Source Software at NASA. *IEEE IT Professional – Special Theme on NASA Contributions to IT, 14*(2), 29-35.

Mattmann, C. A., Waliser, D., Kim, J., Goodale, C., Hart, A., Ramirez, P., et al. (2013). Cloud computing and virtualization within the regional climate model and evaluation system. *Earth Science Informatics*, 1-12.

Miller, R. A., & Frank, W. M. (1993). Radiative forcing of simulated tropical cloud clusters. *Monthly Weather Review, 121*, 482-498.

Miller, D., & Fritsch, J. M. (1991). Mesoscale Convective Complexes in the Western Pacific Region. *Monthly Weather Review , 119*(12), 2978-2992.

Mugnai, A., Smith, E. A., & Tripoli, G. J. (1993). Foundations for Statistical-Physical Precipitation Retrieval from Passive Microwave Satellite Measurements. Part II: Emission-Source and Generalized Weighting-function Properties of a Time-Dependent Cloud-Radiation Model. *Journal of Applied Meteorology, 32*, 17-39.

Mukherjee, D. P., & Acton S. T. (2002). Cloud Tracking by Scale Space Classification. *IEEE Transactions on Geoscience and Remote Sensing, 40*(2), 405–415.

Murakami, M. (1979). Large-Scale Aspects of Deep Convective Activity over the GATE Area. *Monthly Weather Review*, *107*, 994–1013. http://dx.doi.org/10.1175/1520-0493(1979)107<0994:LSAODC>2.0.CO;2

Nalli, N. R., & Reynolds, R. W. (2006). Sea-surface Temperature Daytime Climate Analyses Derived from Aerosol Bias-Corrected Satellite Data. *Journal of Climate, 19*, 410-428.

Nalli, N. R., & Stowe, L. L. (2002). Aerosol correction for remotely sensed sea surface temperatures from the National Ocean and Atmospheric Administration Advanced Very High Resolution Radiometer. *Journal of Geophysical Research, 107*(C10), 3172-3190.

Nesbitt, S. W., Cifelli, R., & Rutledge, S. A. (2006). Storm Morphology and Rainfall

Characteristics of TRMM Precipitation Features. *Montly Weather Review, 134*, 2902-2721.

Nicholson, S. E. (1980). The Nature of Rainfall Fluctuation in Subtropical West Africa. *Monthly Weather Review, 108*, 473-487.

Nicholson, S. E., & Palao, I. M. (1993). A Re-evaluation of Rainfall Variability in the Sahel. Part I. Characteristics of Rainfall Fluctuations. *International Journal of Climatology, 13*, 371-389.

Nicholson, S. E., Some, B., McCollum, J., Nelkin, E., Klotter, D., Berte, Y., et al. (2003a). Validation of TRMM and Other Rainfall Estimates with a High-Density Gauge Dataset for West Africa. Part I: Validataion of GPCC Rainfall Product and Pre-TRMM Satellite and Blended Products. *Journal of Applied Meteorology, 42*, 1337-1354.

Nicholson, S. E., Some, B., McCollum, J., Nelkin, E., Klotter, D., Berte, Y., et al. (2003b). Validation of TRMM and Other Rainfall Estimates with a High-Density Gauge Dataset for West Africa. Part II: Validation of TRMM Rainfall Products. *Journal of Applied Meteorology, 42*, 1355-1368.

Nukulin, G., Jones, C., Fiorgi, F., Asrar, G., Büchner, M., Cerezo-Mota, R., et al. (2012). Precipitation Climatology in an Ensemble of CORDEX-Africa Regional Climate Simulations. *Journal of Climate, 25*, 6057-6078.

Oliphant, T.E. (2007). Python for Scientific Computing. *Computing in Science & Engineering, 9(3)*, 10 – 20.

Otterman, J. (1974). Baring High-Albedo Soils by Overgrazing: A Hypothesized Desertification Mechanism. *Science, 186*, 531-533.

Overpeck, J. T., Meehl, F. A., Bony, S., & Easterling, D. R. (2001). Climate Data Challenges in the 21st Century. *Science, 221*(6018), 700-702.

Parker, D. J., Thorncroft, C. D., Burton, R. R., & Diongue-Niang, A. (2005). Analysis of the African easterly jet, using aircraft observations from the JET2000 experiment. *Quarterly Journal of the Royal Meteorological Society, 131*, 1461-1482.

Pytharoulis, I., & Thorncroft, C. (1999). The Low-level structure of African Easterly Waves in 1995. *Monthly Weather Review, 127*, 2266-2280.

Ray, P. S. (1986). *Mesoscale Meteorology and Forecasting.* (Ed.). Boston, MA: American Meteorological Society.

Redelsperger, J., Thorncroft, C. D., Diedhiou, A., Lebel, T., Parker, D. J., & Polcher, J. (2006). African Monsoon Multidisciplinary Analysis: An international research project and field campaign. *Bulletin of the American Meteorological Society, 87*(12), 1739-1746.

Riehl, H. (1954). *Tropical Meteorology.* New York, NY: McGraw-Hill.

Schröder, M., König, M., & Schmetz, J. (2009). Deep convection observed by the Spinning Enhanced Visible and Infrared Imager on board Meteosat 8: Spatial distribution and temporal evolution over Africa in summer and winter 2006. *Jounral of Geophysical Research Atmospheres, 114*. doi:10.1029/2008JD010653

Semazzi, F. H., & Sun, L. (1997). The role of orography in determining the Sahelian climate. *International Journal of Climatology, 17*, 581-596.

Slingo, J. M. (1987). The Development and Verification of A Cloud Prediction Scheme for the ECMWF model. *Quarterly Journal of the Royal Meteorological Society, 113*(477), 899-927.

Smith, E. A., Mugnai, A., Cooper, H. J., Tripoli, G. J., & Xiang, X. (1992). Foundations for statisical-physical precipitation retrieval from passive microwave satellie measurements. Part I: Brightness-temperature properties of atime-dependent cloud radiation model. *Journal of Applied Meteorology, 31*, 506-531.

Stephens, G. L. (1990). On the Relationship between Water Vapor over the Oceans and Sea Temperature. *Journal of Climate, 3*, 634-645.

Taylor, C. M., Lambin, E. F., Stephenne, N., Harding, R. J., & Essery, R. L. (2002). The Influence of Land Use Change on Climate in the Sahel. *Journal of Climate, 15*, 3615-3629.

Thorncroft, C. D., & Blackburn, M. (1999). Maintenance of the African Easterly Jet. *Quarterly Journal of the Royal Meteorological Society, 125*, 763-786.

Tollerud, E. I., & Collander, R. S. (1993). Mesoscale convective systems and extreme rainfall in the central United States. *Extreme Hydrological Events: Precipitation, Floods and Droughts (Proceedings of the Yokahama Symposium, July, 213.* IAHS Publications.

Trudeau, R. J. (1993). *Introduction to Graph Theory.* Mineola, NY: Dover Publications.

Velasco, I., & Fritsch, J. M. (1987). Mesoscale convective complexes in the Americas. *Journal of Geophysical Research, 92*, 9591-9613.

Vila, D. A., Machado, L. A. T., Laurent, H., & Velasco, I. (2008). Forecast and Tracking the Evolution of Cloud Clusters (ForTraCC) using satellite infrared imagery: Methodology and validation. *Weather & Forecasting*, *23*(2).

Viltard, A., Laurent, H., & De Felice, P. (1990). Lower-Tropospheric Cyclonic Vortices in Nothern Africa in Summer. *Monthly Weather Review, 118*(4), 818-823.

Waliser, D. E., Moncrieff, M. W., Burridge, D., Fink, A. H., Gochis, D., Goswami, B. N., et al.

(2012). The "Year" of Tropical Convection (May 2008 - April 2010): Climate Variability and Weather Highlights. *Bulletin of the American Meteorological Society, 93*(8), 1189-1218.

Webster, P. J., & Stephens, G. L. (1980). Tropical upper-tropospheric extended clouds: Inferences from Winter MONEX. *Journal of Atmospheric Sciences, 37*, 1521-1541.

West Africa: Facts and Figures. (n.d.). Retrieved from http://web.worldbank.org/WBSITE/EXTERNAL/NEWS/0,,contentMDK:20179737~pagePK:34370~piPK:42768~theSitePK:4607,00.html.

West Africa floods: Situation report No. 2. (2009, September 29). *World Health Organization.* Retrieved from http://reliefweb.int/report/burkina-faso/west-africa-floods-situation-report-no-2.

White, T. (2012). *Hadoop: The definitive guide*. " O'Reilly Media, Inc.".

Whitehall, K., Chiao, S., & Mayers-Als, M. (2013). Numerical Investigations of Convective Initiation in Barbados. *Advances in Meteorology.* doi:10.1155/2013/630263

Whitehall, K., Mattmann, C., Waliser, D., Kim, J., Goodale, C., Hart, A., et al. (2012). Building Model Evaluation and Decision Support Capacity for CORDEX. *WMO Bulletin, 61*(2), 29-34.

Williams, M., Houze Jr., R. A. (1987). Satellite-observed characteristics of winter monsoon cloud clusters. *Monthly Weather Review, 115*(22), 505–519.

Woodley, W. L., Griffith, C. G., Griffin, J. S., & Stromatt, S. C. (1980). The interference of GATE convective rainfall from SMS-1 imagery. *Journal of Applied Meteorology, 19*(4), 338-408.

Xu, R., & Wunsch, II, D. C. (2008). *Clustering*. Hoboken, NJ: Wiley-IEEE Press.

Xue, Y., & Shukla, J. (1993). The influence of land surface properties on Sahel climate. Part 1: Desertification. *Journal of climate*, *6*(12), 2232-2245.

Xue, Y., & Shukla, J. (1996). The influence of land surface properties on Sahel climate. Part II: Afforestation. *Journal of Climate*, *9*, 3260–3275. http://dx.doi.org/10.1175/1520-0442(1996)009<3260:TIOLSP>2.0.CO;2

Zeng, N., Neelin, J. D., Lau, K. M., & Tucker, C. J. (1999). Enhancement of Interdecadal Climate Variability in the Sahel by Vegetation Interaction. *Science, 286*, 1537-1540.