A STATISTICAL COMPARISON OF FOUR PRECIPITATION CLIMATE DATASETS
IN THE GANGES-BRAHMAPUTRA-MEGHNA RIVER BASIN

by

JAMES BRENTON

A THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science
in
The Earth System Science Program
to
The School of Graduate Studies
of
The University of Alabama in Huntsville

HUNTSVILLE, ALABAMA

2014
In presenting this thesis in partial fulfillment of the requirements for a master's degree from The University of Alabama in Huntsville, I agree that the Library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by my advisor or, in his/her absence, by the Chair of the Department of the Dean of the School of Graduate Studies. It is also understood that due recognition shall be given to me and to the The University of Alabama in Huntsville in any scholarly use which may be made of any material in this thesis.

James Brenton

June 27, 2014
THESIS APPROVAL FORM

Submitted by James Brenton in partial fulfillment of the requirements for the degree of Master of Science in Earth System Science and accepted on behalf of the Faculty of the School of Graduate Studies by the thesis committee.

We, the undersigned members of the Graduate Faculty of The University of Alabama in Huntsville, certify that we have advised and/or supervised the candidate on the work described in this thesis. We further certify that we have reviewed the thesis manuscript and approve it in partial fulfillment of the requirements for the degree of Master of Science in Earth System Science.

Dr. Rob Griffin (Date) 7/23/2014 Committee Chair

Dr. Pete Robertson 8/8/2014

Dr. Udaysankar Nair 8/8/2014 Interim Dept. Chair

Dr. Larry Carey 8/12/2014 Dean of College of Science

Dr. Sundar Christopher 8/12/2014 Dean of Graduate Studies

Dr. David Berkowitz
ABSTRACT
The School of Graduate Studies
The University of Alabama in Huntsville

Degree: Master of Science Program: Earth System Science
Name of Candidate: James Brenton

There are 630 million people that live within the Ganges-Barhmaputra-Meghna River basin (GBM) that depend on these rivers for their livelihood (Frenken, 2011). These rivers greatly depend upon the precipitation and rain fall to contribute to their flows. Researchers and endusers in the GBM basin depend on climate precipitation datasets to help them study the impact precipitation has on hydrology in this region. To properly utilize a dataset, the source input data, the algorithms used to merge the data, and the final temporal and spatial resolution of the dataset all need to be properly understood. The goal for this study was to statistically compare four climate precipitation datasets, GPCP, CMAP, CHIRPS and APHRODITE, in the GBM basin to identify significant differences among them. Using an empirical distribution function, Pearson correlation, Spearman’s Rank correlation, and Anomaly correlation, this study showed that across all elevations and during the monsoon and dry seasons, CMAP disagreed the most from the other datasets, which comes from the difference in input source data and the merging algorithm of CMAP and the other datasets. This study also found that GPCP and APHRODITE were in the highest agreement with the other datasets at a monthly temporal resolution during the dry seasons. Also, APHRODITE is reporting different rainfall rates in the mountainous regions during the monsoon season than the other datasets, implying that the rain gauges used by APHRODITE are reporting a different rate of precipitation than the remotely sensed satellite precipitation products in the mountainous region. These recommendations and data will be shared with the SERVIR science team and the hydrologists at ICIMOD so they know how the datasets differ in their descriptions of climatic precipitation over the GBM basin.

Abstract Approval: Committee Chair
Dr. Robert Griffin

Department Chair
Dr. Larry Carey

Graduate Dean:
Dr. David Berkowitz
ACKNOWLEDGEMENTS

The work presented in this thesis was made possible with the support, assistance, and guidance from several people including my committee chair, Dr. Robert Griffin, for his academic and professional guidance and technical assistance, my other committee members, Dr. Pete Robertson, for his advice and assistance with data analysis, and Dr. Udaysankar Nair for his assistance with climate analysis. I’d also like to thank Dr. Brent Roberts of the SERVIR science team for his guidance and assistance for interpreting the statistical methods used in this thesis. Without the help of these four professors and scientists, I could not have completed this research, and I’m very grateful for the time and effort they invested into me. Special thanks also go to NASA’s SERVIR program and Dan Irwin for their support during my time as a graduate student at The University of Alabama in Huntsville.

This thesis would also not be possible without the personal support of my family, specifically, my parents, Chuck and Terri Brenton, and my brother, Jacob Brenton. Throughout my time at The University of Alabama in Huntsville they always challenged and supported me to reach my goals.

Finally, I’d like to thank my wife, Elise Brenton. For the late nights I kept her up while I was working, for all the trips to the lab over the weekend, and for the constant love and support, I could not have done this without her.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>List of Figures</th>
<th>viii</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>List of Acronyms</td>
<td>xi</td>
</tr>
</tbody>
</table>

## Chapter

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>A. Purpose of Study</td>
<td>1</td>
</tr>
<tr>
<td>B. Specific Objectives</td>
<td>2</td>
</tr>
<tr>
<td>C. Summary</td>
<td>4</td>
</tr>
<tr>
<td>2. BACKGROUND AND STUDY AREA</td>
<td>5</td>
</tr>
<tr>
<td>A. Climate</td>
<td>5</td>
</tr>
<tr>
<td>B. Hydrology</td>
<td>8</td>
</tr>
<tr>
<td>C. Social Needs</td>
<td>10</td>
</tr>
<tr>
<td>3. LITERATURE REVIEW</td>
<td>12</td>
</tr>
<tr>
<td>A. Measuring Precipitation</td>
<td>12</td>
</tr>
<tr>
<td>B. Merged Precipitation Datasets</td>
<td>17</td>
</tr>
<tr>
<td>C. Climate Studies in the Ganges-Brahmaputra-Meghna River Basin</td>
<td>32</td>
</tr>
<tr>
<td>D. Needs from the Scientific Community</td>
<td>35</td>
</tr>
<tr>
<td>4. METHODOLOGY</td>
<td>37</td>
</tr>
<tr>
<td>A. Data Preprocessing</td>
<td>37</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Name</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Map of Ganges-Brahmaputra-Meghna River basin</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>Monsoon rainfall totals across India and the GBM basin</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>Rivers and tributaries for each of the three subbasins</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>Atmospheric Transmission Windows</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>GPCP Merging Technique</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>CMAP Merging Technique</td>
<td>25</td>
</tr>
<tr>
<td>7</td>
<td>APHRODITE Merging Technique</td>
<td>28</td>
</tr>
<tr>
<td>8</td>
<td>CHIRP Merging Technique</td>
<td>31</td>
</tr>
<tr>
<td>9</td>
<td>CHIRPS Merging Technique</td>
<td>32</td>
</tr>
<tr>
<td>10</td>
<td>DEM of GBM Basin and study grid</td>
<td>43</td>
</tr>
<tr>
<td>11</td>
<td>Elevation classification</td>
<td>44</td>
</tr>
<tr>
<td>12</td>
<td>APHRODITE average station density</td>
<td>46</td>
</tr>
<tr>
<td>13</td>
<td>CHIRPS average station density</td>
<td>47</td>
</tr>
<tr>
<td>14</td>
<td>GPCC average station density</td>
<td>47</td>
</tr>
<tr>
<td>15</td>
<td>Empirical Distribution Function of negative anomalies</td>
<td>58</td>
</tr>
<tr>
<td>16</td>
<td>Empirical Distribution Function of positive anomalies</td>
<td>60</td>
</tr>
<tr>
<td>17</td>
<td>Correlation Maps for the entire study period</td>
<td>61</td>
</tr>
<tr>
<td>18</td>
<td>Annual correlations of anomalies</td>
<td>63</td>
</tr>
<tr>
<td>19</td>
<td>Seasonal Correlations between all datasets</td>
<td>63</td>
</tr>
<tr>
<td>20</td>
<td>Low, Dry Season GPCP-APHRODITE Anomaly Correlation and Rain Gauge Stations</td>
<td>69</td>
</tr>
<tr>
<td>21</td>
<td>Low, Dry Season CMAP-APHRODITE Anomaly Correlation and Rain Gauge Stations</td>
<td>69</td>
</tr>
<tr>
<td>22</td>
<td>Low Lands, Monsoon Season enhanced CMAP-GPCP Anomaly Correlation with Rain Gauge Station</td>
<td>71</td>
</tr>
<tr>
<td>23</td>
<td>Low Lands, Monsoon Season, standard CMAP - GPCP Anomaly Correlation with Average Rain Gauge Stations</td>
<td>71</td>
</tr>
<tr>
<td>24</td>
<td>CMAP-enhanced correlated with GPCP, with GPCC rain gauge density for transition elevations during the dry season</td>
<td>72</td>
</tr>
<tr>
<td>25</td>
<td>CMAP-standard correlated with GPCP, with GPCC rain gauge density for transition elevations during the dry season</td>
<td>72</td>
</tr>
<tr>
<td>26</td>
<td>CMAP-standard correlated with GPCP, with GPCC rain gauge density for lower elevations during the monsoon season</td>
<td>73</td>
</tr>
<tr>
<td>27</td>
<td>CMAP-enhanced correlated with CHIRPS, with GPCC and CHIRPS rain gauge density for mountainous elevations during the monsoon season</td>
<td>74</td>
</tr>
<tr>
<td>28</td>
<td>CMAP-enhanced correlated with APHRODITE, with GPCC and APHRODITE rain gauge density for lower elevations during the monsoon season</td>
<td>75</td>
</tr>
<tr>
<td>29</td>
<td>Low Lands, Dry Season CMAP-CHIRPS Anomaly Correlation with Average Rain Gauges per cell</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>Description</td>
<td>Correlation</td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>30</td>
<td>Mountainous Land, Dry Season CMAP-CHIRPS Anomaly Correlation with Average Rain Gauges per grid cell</td>
<td>76</td>
</tr>
<tr>
<td>31</td>
<td>Mountainous, Dry Season GPCP-APHRODITE Anomaly Correlation with Average Rain Gauges per grid cell</td>
<td>78</td>
</tr>
<tr>
<td>32</td>
<td>Low Lands, Dry Season Annual Anomaly Correlation</td>
<td>79</td>
</tr>
<tr>
<td>33</td>
<td>Transitional Elevations, Dry Season Annual Anomaly Correlation</td>
<td>80</td>
</tr>
<tr>
<td>34</td>
<td>Mountainous, Dry Season Annual Anomaly Correlation</td>
<td>81</td>
</tr>
<tr>
<td>35</td>
<td>Low Lands, Monsoon Season Annual Anomaly Correlation</td>
<td>82</td>
</tr>
<tr>
<td>36</td>
<td>Transitional Elevations, Monsoon Season Annual Anomaly Correlation</td>
<td>83</td>
</tr>
<tr>
<td>37</td>
<td>Mountainous, Monsoon Season Annual Anomaly Correlation</td>
<td>84</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Merged Datasets descriptions</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>ESRI ASCII Grid text format</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>Spearman’s Rank Correlation, Monsoon season</td>
<td>67</td>
</tr>
<tr>
<td>4</td>
<td>Spearman’s Rank Correlation, Dry season</td>
<td>67</td>
</tr>
</tbody>
</table>
## LIST OF ACRONYMS

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition of Acronym</th>
</tr>
</thead>
<tbody>
<tr>
<td>APHRODITE</td>
<td>Asian Precipitation - Highly-Resolved Observational Data Integration Towards Evaluation of Water Resources</td>
</tr>
<tr>
<td>CDIAC</td>
<td>Carbon Dioxide Information Analysis center</td>
</tr>
<tr>
<td>CHIRPS</td>
<td>Climate Hazard Group InfraRed Precipitation with Station</td>
</tr>
<tr>
<td>CMAP</td>
<td>Climate Prediction Center Merged Analysis Project</td>
</tr>
<tr>
<td>ECAD</td>
<td>European climate assessment and Dataset</td>
</tr>
<tr>
<td>FAO</td>
<td>Food and Agriculture Organization</td>
</tr>
<tr>
<td>GAME-T</td>
<td>Global Energy and Water Cycle Experiment Asian Monsoon Experiment in Tropics</td>
</tr>
<tr>
<td>GBM</td>
<td>Ganges-Brahmaputra-Meghna River Basin</td>
</tr>
<tr>
<td>GHCN</td>
<td>Global Historical Climatology Network</td>
</tr>
<tr>
<td>GPCC</td>
<td>Global Precipitation Climatology Centre</td>
</tr>
<tr>
<td>GPCP</td>
<td>Global Precipitation Climate Project</td>
</tr>
<tr>
<td>GPI</td>
<td>Geostationary Operational Environmental Satellite Precipitation Index</td>
</tr>
<tr>
<td>MRC</td>
<td>Mekong River Commission</td>
</tr>
<tr>
<td>MSU</td>
<td>Microwave Sounding Unit</td>
</tr>
<tr>
<td>NCAR-DS</td>
<td>National Center for Atmospheric Research Data Archive</td>
</tr>
<tr>
<td>NCDC</td>
<td>National Climatic Data Center</td>
</tr>
<tr>
<td>NCEP/NCAR</td>
<td>National Centers for Environmental Prediction/National Center for Atmospheric Research</td>
</tr>
<tr>
<td>OPI</td>
<td>Outgoing Longwave Radiation Precipitation Index</td>
</tr>
<tr>
<td>SSM/I</td>
<td>Special Sensor Microwave Imager</td>
</tr>
<tr>
<td>SSMIS</td>
<td>Special Sensor Microwave Imager Sounder</td>
</tr>
<tr>
<td>TOVS</td>
<td>Television Infrared Observation Satellite Operational Vertical Sounder</td>
</tr>
<tr>
<td>TRMM</td>
<td>Tropical Rainfall Measuring Mission</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

A. Purpose of Study

In late June 2012, Chittagong, Bangladesh, a city of 6.5 million people, experienced a torrential downpour of rain. While it is normal for the monsoon rain to fall during this time of year, this was heavier than usual. Bangladesh officials reported that this event was one of the heaviest monsoon rain events in many years for their region. The flooding began as rain water fell and drained into the Karnphuli river basin. The river began to swell and it overcame the banks, flooding the streets and the tin roof homes of Chittagong. Making the situation worse, Chittagong is the second largest city in Bangladesh and excessive flooding displaced thousands, stranding them on whatever high ground they could find. When the rains had stopped, 110 people were killed in the flooding and subsequent landslides (BBC, 2012).

The June 2012 flood is not the first major flood, and it will not be the last. In 1998, over 75% of Bangladesh was flooded due to monsoon rainfall coinciding with peak flows of the
Ganges, Brahmaputra, and Meghna rivers (BBC, 1998). The rainfall and flooding is not just limited to just Bangladesh, as major flooding has recently occurred in Northern India and Nepal in June 2013 killing hundreds and causing millions of dollars’ worth of damage (ICIMOD, 2013). Clearly, there is a need to understand the climate and impact of the monsoon in the basin for the Ganges, Brahmaputra, and Meghna Rivers in order to protect lives and property. This need requires accurate precipitation amounts from satellites and gauges that are at a fine scale to drive agricultural and hydrologic models.

The goal of this study is to generate a measure of correlation among climatic precipitation datasets for the GBM river basin. This study will review some of the most commonly used and studied climatic precipitation datasets, the Global Precipitation Climate Project (GPCP) and the Climate Prediction Center Merged Analysis of Precipitation (CMAP), and some newer datasets that offer higher spatial resolution, Climate Hazards Group Infrared Precipitation with Stations (CHIRPS) and the Asian Precipitation – Highly-Resolved Observational Data Integration Towards Evaluation of Water Resources (APHRODITE). Researchers and decision makers need to know where datasets agree and disagree before they use it for monitoring, forecasting, or modeling precipitation for such uses as hydrology and agriculture. These findings and recommendations will provide the end-users with an informed conclusion on the agreement between datasets.

B. Specific Objectives

This study will address this goal through four objectives and use the four datasets: CMAP, GPCP, APHRODITE, CHIRPS:
1. To find a broad temporal and spatial metric to determine which datasets are reporting
greater or lesser precipitation over the entire study area and during the entire study period.

2. To find a measurement that will describe agreement and disagreement at different
elevation zones. Topography plays a major role in the GBM basin’s interaction with the
monsoon. Because of this, it is important to observe how much rain falls in regions of similar
topography,

3. To find a measurement that will describe agreement and disagreement during the dry and
monsoon season.

4. To determine if the density of rain gauges in the development of datasets have affected
the correlation between the datasets.

The thesis will address this goal and these objectives through the discussions provided in
each chapter. Chapter 1: Introduction will outline the goal and the objectives. In Chapter 2:
Background and Study Area, there will be a discussion on the formation and impact of the
monsoon on the GBM basin, which will provide the background information needed in the other
chapters to meet the objectives. Then, there will be a discussion on the impact of monsoon
precipitation on the hydrology of the GBM basin. Finally, the chapter will conclude with a
discussion of the social needs for the precipitation data. In Chapter 3: Literature Review, there
will be a discussion that outlines the methods of recording precipitation using rain gauges and
satellites. This chapter will have another discussion that will outline how each of the four
datasets merges rain gauge and/or satellite data, and another discussion on relevant climate
precipitation studies in the GBM basin. Finally, this chapter will conclude with a discussion on
the scientific community’s need for this comparative study of climatic precipitation data. Similar
to Chapter 2, Chapter 3 provides necessary information to be used in chapters four, five, and six. In Chapter 4: Data and Methodology, the preprocessing of the data will be outlined so that it will be clear how the data was obtained and prepared. Then, there will be a discussion on how the Digital Elevation Model (DEM) data was obtained and used, followed by a discussion of how a common time period was selected for the study based on the temporal resolution of the data. Next, there will be discussion on the rain gauge station locations and their density for each dataset. Then, Chapter 4 will discuss the various statistical methods that will be used and how each method will address one of the four objectives. In Chapter 5: Findings, the results from the statistical tests described in Chapter 4 will be discussed. This will further answer how these statistical tests and values meet the needs set forth by the objectives. In Chapter 6: Discussion, the implications and suggestions from the findings will be discussed, and finally, the goal of the study will be discussed.

C. Summary

The purpose of this study is to determine when and where GPCP, CMAP, CHIRPS, and APHRODITE differ significantly. This will have a major impact on researchers that are looking to use this data in agricultural or hydrology studies, as these differences will be reflected in their results, potentially affecting the people of the GBM basin. Furthermore, this study will be useful to determine if the newer datasets, CHIRPS and APHRODITE, are in agreement with the well-established datasets, GPCP and CMAP, thus showing that these higher resolution datasets are viable alternatives for climate studies in this region.
CHAPTER 2

BACKGROUND AND STUDY AREA

A. Climate

The Ganges-Brahmaputra-Meghna River basin’s climate, hydrology, and even the lives of the people native to the basin are widely defined by the monsoon. However, depending on the specific location within the basin, there can be widely varied climatologies; from the drier northwest upper region of the Ganges river to the high precipitation near the coast where the Ganges meets the Brahmaputra and the Meghna Rivers just a few hundred kilometers upstream from the mouth of the Bay of Bengal (Frenken, 2011). These main three rivers and many of their tributaries are transboundary, making their water essential for many people across the basin, while adding a degree of difficulty to secure and protect water for people due to multiple government entities. The GBM river system is the third largest freshwater outlet to the world’s oceans, followed by the Amazon and Congo River systems (Chowdhury and Ward, 2004).
There have been several studies that look to link the start of the monsoon to atmospheric and oceanic characteristics in the Indian Ocean (Shrestha, 2000, Lang and Barros 2001, Lang and Barros 2002). These studies have shown that the onset of the monsoon over Nepal and Northern India is caused by an early depression in the Bay of Bengal which generally occurs in the month of May, ahead of the onset of the monsoon in June. Also, the amplitude of the monsoon is linked to the Southern Oscillation Index of the sea surface temperature (SST) of the Indian Ocean. In other words, warm SST anomalies in the Indian Ocean near the Bay of Bengal typically result in excess monsoon rainfall, while cooler SST anomalies in the Bay of Bengal result in a deficient monsoon rainfall. Taking these findings into account, it can be said that in order for the monsoon to form and provide large amounts of rainfall, the atmosphere needs to
have a low pressure circulation form over warm SST anomalies, where water vapor is evaporated off the ocean surface, and condensed into the clouds that will make their way up the coast of the Indian Sub-continent (Fig. 2).

Figure 2: Monsoon rainfall totals across India and the GBM basin (Shulbuch, 2010)

Once these monsoon rain clouds have made landfall, they continue to move north until they reach the barrier of the Himalayan Mountains. Barros and Lang (2001) outlined several differences in the monsoon across the diurnal cycle. This study found that the mountains block the diurnal flows of wind, causing convergence at low levels. During the day, the convergence is reduced due to diurnally forced upslope flow reducing the spatial gradient in the wind velocities.
Furthermore, the upslope flow is responsible for convection at higher elevations, resulting in more precipitation during the day at higher altitudes. During the night, the weaker surface wind speeds do not have as strong upslope flow, and thus converge at lower elevations, finally forcing convection and more rainfall in the lower elevations at night than during the day.

The monsoon in the GBM basin delivers 80% of the annual rainfall in the region. (Frenken, 2011). Given such a large contribution, there is a need to be able to monitor past monsoon events and to use this climate data to create models that could predict future monsoons. Future climate change predictions have massive uncertainties because of the importance of rainfall and the lack of data in mountainous and low populated areas (Jeuland et al, 2013). The IPCC Fourth Assessment Report warned that South Asia and the greater Himalayans have widely divergent predictions of future changes in precipitation (IPCC, 2007). In the recent study conducted by Jeuland et al. (2013) the authors preface their research by saying there is an underlying lack of confidence in their results in hydrological impacts from climate prediction which stems from a lack of data. APHRODITE is one of the few datasets that offer directly observed data from beyond 30 years into the past. APHRODITE offers data from 1951 to 2007, but it is useful to compare it to other datasets that use satellite data. Based on these comparisons, it can be determined how well the APHRODITE dataset agrees with other precipitation datasets which merge remotely sensed multispectral information whose data only covers the last 30 years.

B. Hydrology

The Ganges, Brahmaputra, and Meghna rivers respond to monsoon rainfall in extreme ways. The Ganges River and its tributaries (Fig. 3) rise from very low levels in May to a peak in the later months of the monsoon. For example, along the southern border between India and
Bangladesh the vast majority of flow of the Ganges River comes in July through October, and only 6% of the total annual flow comes during the dry winter and spring months, January through May (Jeuland, et al. 2013). These values are very susceptible to variation in precipitation can greatly changes these numbers if a dry or monsoon season lasts longer than usual. During short periods of heavy rainfall, large amounts of silt are moved downstream (ESA, 2011). In fact, over one billion tons of silt are moved every year by the Ganges and Brahmaputra River (Kuehl et al., 2011). During periods of high precipitation and high flows, crop production is increased during the winter planting season (Yu et al, 2010) even though the silt carried by the rivers is not particularly fertile (Subramanian et al, 1996).

It is also important to note that only 2% of the total river flow comes from glacial melt (Jeuland et al. 2013). This is very different from other rivers around the world. Even the Indus River, the closest major river basin, receives the majority of its flow from glacial melt. Most of the glacial melt that contributes to the flow of the Ganges occurs during the early monsoon (Alford and Armstrong, 2010), yet despite its minor contribution, glacial melt is key to maintaining the perennial flow of the Himalayan tributaries of the Ganges. Snowmelt likely

*Figure 3. Rivers and tributaries for each of the three subbasins. (Kjelds and Jørgensen, 1997)*
contributes more to the flow within the basin than glacial melt, but this is difficult to determine since there is not a large number of precise estimates of snowmelt (Jeuland et al. 2013).

In the low-lying plains of the GBM basin, 90% of all available potable water is used for agricultural purposes (World Bank, 2013), yet not all of this comes from surface water. The surface and groundwater irrigation systems in the Ganges plain are very dependent on one another: the Basin’s need for water is supplemented by pumping water from aquifers and other ground water sources (Shah, 2008), and this surface irrigation often seeps into the ground water, recharging the underground water storage. In contrast to the widely available stored groundwater, there is very little stored surface water because many of the dams and reservoirs in this region along the Ganges River are relatively small (all but five are taller than 100 meters) (Jeuland et al. 2013), meaning that the amount of precipitation that falls during the monsoon season will have a major effect on the flow in the rivers of the GBM basin.

C Social Needs

Given the extreme variability of flow as a result of the torrential monsoon rainfall, the estimated 630 million people that live in the GBM basin must learn how to cope with this rainfall and the subsequent flooding, which may result in the loss of lives, homes, crops, and livestock (Ahmad, 2007).

The GBM basin is comprised of a total surface area of 1.7 million square kilometers including parts of India (64% of total surface area), China (18% of total surface area), the entire country of Nepal (9% of total surface area), Bangladesh (7% of total surface area), and the entire country of Bhutan (3% of total surface area). Given the varied topography and climate within each of these countries, there is a wide range in terms of population density across the basin. For
example, in the mountainous regions of China and Bhutan population density ranges from 6 to 18 people per square kilometer, and from 195 to 1,013 people per square kilometer in Nepal, India, and Bangladesh respectively. The GBM river basin is home to the largest number of the world’s poor in any one region (Frenken, 2011). The population density is increasing across the basin. Furthermore, the population density has already shown to be at high levels in India and Bangladesh. The most populous countries in the region, Nepal, India, and Bangladesh, have had difficulties integrating a water resource and development plan in the past (Biswas and Uitto, 2001) and before decision makers can develop plans, a proper understanding of where and how much water is available from precipitation is needed. It is understood that each precipitation dataset will have its biases and errors, so it is important that the difference between the datasets be understood before researchers and decision makers begin developing models and policy based on these data, thus the goals of this study.
CHAPTER 3

LITERATURE REVIEW

A. Measuring Precipitation

Rain gauges have come far in the last several hundred years; their design has been modified to more accurately record rainfall, but there are still uncertainties in their records. For instance, rain gauges still need to protect against losses due to evaporation, losses due to wetting of the gauges, over-measurement due to splash from the surrounding area, placement of rain gauge, under-measurement due to turbulence around the gauge. However, even with these improvements in rain gauge design, there is still the issue of spatial distribution. It can be argued that even if a rain gauge has an accurate design, the rain gauge is only accurate for the point at which is placed (Davie, 2008).

Because of the uneven distribution of rain gauges across the globe and the high demand for rain gauge data, there are numerous studies that have explained various interpolation methods that can be used to create an estimated spatial distribution of precipitation. Some of the most
common interpolation methods include the Inverse Distance Weighted average (IDW), Splining, and Kriging. IDW is a simple, but efficient interpolation method that illustrates the effects of distance interpretation in interpolation. While there are various types of IDW (such as beta and gamma), the standard IDW interpolation is estimated by a weighted mean of the observations (Ahrens, 2006). The weighted values are proportional to a negative power of geographical distances between the point of interpolation and the observation, or rain gauge. This is shown in this equation:

$$P_0 = \frac{\sum_{a=1}^{n} P_a \omega_\alpha}{\sum_{a=1}^{n} \omega_\alpha}, \text{ where } \omega_\alpha = \frac{1}{d_\alpha^\lambda}.$$  

Here $P_0$ is the value of the cell being interpolated, $P_a$ is all of the observations, $d_\alpha$ is the geographical distance between the point of interpolation and the nearest observation, or rain gauge, and the $\lambda$ power of distance must be carefully selected depending on the interpolated variable. A value of two for $\lambda$ is often used for these calculations. If $n=1$ then only the next neighbor of the point being interpolated is considered, giving the same results as the Thiessen method (Ahrens, 2006). The Spline interpolation technique is different from the IDW in that this function “smooths” the values being interpolated rather than base the new values on the neighboring interpolated points. This “smoothing” represents the minimizing of the deviation from observations, or rain gauges. A Spline interpolation, $S(x)$, is represented by:

$$S(x) = T(x) + \sum_{j=1}^{N} \lambda_j R(x, x^{(j)})$$

In this equation $T(x)$ is a trend function and $R(x, x^{(j)})$ is a radial basis function which is dependent on the choice for the smoothness seminorm. While there are various other splining
techniques, such as thin plate spline, all splining techniques will possess a trend function and a radial basis function. The Kriging method is different from IDW and Splining in that it uses a more statistical approach to interpolation. Kriging is a generalized least-square regression technique that accounts for the spatial dependence between observations and makes spatial predictions of values (Goovaerts, 2000). The amount, $z_{OK}^*(u)$, of precipitation predicted by Kriging can be approximated by:

$$z_{OK}^*(u) = \sum_{a=1}^{n(u)} \lambda_{a}^{OK}(u) z(u_a) \text{ where } \sum_{a=1}^{n(u)} \lambda_{a}^{OK}(u) = 1$$

Here, an observation, or rain gauge measurement, is represented by $z(u_a)$ and $\lambda_{a}^{OK}(u)$ is the ordinary kriging weights. There are three common methods for determining these kriging weights: spherical, cubic, and dampered whole effect model. Out of these three, the spherical model is most widely used, because the weighting function has a linear behavior that is easier to calculate and draw direct “cause and effect” conclusions (Goovaerts, 2000).

Obviously, these are only the most basic interpolation descriptions, and many other researchers have improved upon these methods to account for surface features, topography, and the presence of large bodies of water. APHRODITE and CHIRPS use their own interpolation techniques that borrow upon similar calculations as IDW and Kriging interpolation (Yatagai et al. 2012, Funk et al. 2013). However, interpolation will always have errors or biases associated with their methods of spatial estimation of precipitation.

A relatively newer alternative is the use of precipitation estimates from satellite based instruments. Satellites can provide measurements of temperatures, humidity, and precipitation because of their ability to interpret radiation across the electromagnetic spectrum that has been absorbed, scattered, or emitted by the atmosphere. The electromagnetic spectrum is the range of
radiation, as measured in wavelength, thus assuming that energy behaves as waves and not particles.

These interactions need to be understood when using satellite remotely sensed data so that the proper atmospheric constituent or surface feature is observed, by the use of atmospheric windows. Atmospheric windows are regions in the electromagnetic spectrum that are less affected by atmospheric scattering and absorption (Fig. 4). These windows occur because of the presence of specific gases in the atmosphere and their respective absorption bands across the electromagnetic spectrum. The chemical composition of each atmospheric constituent, such as ozone, oxygen, water vapor, and carbon dioxide, determines the absorption and transmission of each wavelength through the visible, IR, and Microwave portions of the electromagnetic spectrum (Richard, 2012). This is especially important for selecting wavelengths to view the surface. But for this study, the presence of atmospheric windows will help to determine which wavelengths of the microwave and IR portions of the electromagnetic spectrum the merged datasets use.

![Graph of Atmospheric Transmission Windows](image)

*Figure 4: Atmospheric Transmission Windows (Liew, 2001)*

Although the errors are different from rain gauges, satellite remotely sensed data has its own challenges for compensating for errors within the data. These errors must be corrected
before the data can be used in the merging process of each dataset. There are two main types of errors associated with using remotely sensed data, radiometric errors and geometric errors.

Radiometric errors occur from issues with the accuracy of surface spectral reflectance, emittance, or back-scattered measurements from remotely sensed instruments. There are three types of radiometric errors: sensor error, atmospheric error, and topographic error. Sensor error occurs when there are random bad pixels (noise), line or column drop-outs, and line or column striping. Noise can be corrected by using threshold algorithms to “smooth” data. Line or column drop-outs can be corrected through simple horizontal adjustments. Line or column striping can be corrected by locating the bad lines in the datasets and assigns values to these pixels based on neighboring pixels. Finally, topographic error occurs when there is radiometric distortion from slope and aspect. This can be correct by removing topographically induced illumination variation so that two objects with the same reflectance properties will have the same brightness value in the image, despite their different orientation to the Sun’s position.

Geometric errors occur from the spatial misrepresentation of remotely sensed data. This could include the scanning system-induced variation in ground resolution cell size, altitude changes affecting the viewing angle. These errors can be corrected by preprocessing or post processing the data for georeferencing or spatial interpolation.

The satellite products used by the merged climatic precipitation datasets use Infrared (IR) data to measure cloud top temperatures. This is measured in the 11 to 12 μm region of the electromagnetic spectrum. This region is specific for satellite remotely sensed IR data because there is an atmospheric window. The cloud top temperature is used to determine cloud top height. Given the height of cloud tops, the likelihood of precipitation can be determined
(Giannakos and Feidas 2012). However, the height of clouds and presence of rain changes with latitude, as well as the presence of non-precipitous cirrus or other clouds make the use of IR data to study precipitation difficult (Giannakos and Feidas 2012).

The merged climate precipitation datasets use microwave data at 19.35, 21.235, 37.0, and 85.5 μm. Unlike the IR data, the microwave data has a much longer wavelength. This longer wavelength is sensitive to microphysical properties instead of cloud-top temperature, as with IR. Microwave radiation can penetrate through clouds and the atmosphere, giving it an advantage over IR data, which can only be used to study cloud tops. The clouds are opaque in the IR region of the spectrum, giving microwave an advantage to studying the macrophysical properties of precipitation.

Raindrops in the atmosphere absorb and re-emit microwave radiation at their own thermodynamic temperature. Furthermore, the longer microwave wavelengths tend to saturate at higher rainfall rates and are less sensitive to the effects of ice scattering (Kummerow et al. 1996). The longer microwave wavelengths are much lower powered then IR, so a wider field of view is necessary to get enough microwave radiation to appear in images. When both IR and microwave data are used, the IR data can be used to find the location of potential rain clouds, and the microwave data can be used to determine the amount of precipitation in the cloud.

B. Merged Precipitation Datasets

The climatic precipitation data sets used in this study were the Global Precipitation Climate Project (GPCP), Climate Prediction Center (CPC) Merged Analysis of Precipitation (CMAP), the Asian Precipitation – Highly Resolved Observational Data Integration Towards
Evaluation (APHRODITE) and the Climate Hazard group with InfraRed with Stations (CHIRPS). These climatic precipitation datasets have unique sources of data and algorithms that they rely on to produce their respective data sets. Before a description of these differences, a shared temporal resolution must be established for comparison. GPCP and CMAP are available from January 1979 to the present (with delay), APHRODITE is available from January 1951 to December 2007, and CHIRPS is available from January 1981 to near present. Because of this timing, the study period will be from January 1981 to December 2007. A table (Table 1) has been provided below to further illustrate the similarities and differences in the spatial and temporal resolutions, as well as the similarities and differences in the input data sources used by each merged dataset.
<table>
<thead>
<tr>
<th>Name</th>
<th>Spatial Resolution</th>
<th>Temporal Resolution</th>
<th>Source Data</th>
</tr>
</thead>
</table>
| GPCP     | 2.5° x 2.5° (Global Coverage) | Jan. 1979 – Present (Monthly) | - SSM/I emission estimates  
- F13 SSMIS: calibrating microwave data source  
- SSM/I scattering estimates  
- GPI and OPI estimates  
- TOVS estimates  
- GPCC rain gauge analyses |
| CMAP     | 2.5° x 2.5° (Global Coverage) | Jan. 1979 – Present (Monthly and Pentad) | - SSM/I emission estimates  
- MSU microwave data  
- SSM/I scattering  
- GPI and OPI estimates  
- NCEP/NCAR Reanalysis Precipitation  
- GPCC rain gauge analyses |
| APHRODITE | 0.5° x 0.5° (Monsoon Asia, 60°E-15°E, 15°S-55°N) | Jan. 1951 – Dec. 2007 (Daily) | Station Data supplied by:  
- Local Meteorological or Hydrological organizations  
- Pre-compiled datasets (GHCN, CDIAC, NCAR-DS, NCDC, FAO, GAME-T, MRC, and ECAD)  
- Global Telecommunication System |
| CHIRPS   | 0.05° x 0.05° (Near Global coverage, 50°S-50°N across all latitudes) | Jan. 1981 – Present (Monthly, Pentad, and Dekad) | - Monthly precipitation climatology  
- Quasi-global geostationary thermal IR satellite observations  
- TRMM 3-hour estimate  
- Atmospheric model rainfall from NOAA Climate Forecast System  
- In-situ precipitation observations from national and regional meteorological services. |

Table 1: Table of spatial and temporal resolution, and source data for each dataset
The goal of GPCP is to provide a collection of monthly and finer temporal resolution analyses with global coverage at 2.5° resolution with 1° for daily data sets. For this study, the monthly, 2.5° data will be used to match the monthly temporal scale of the CHIRPS data. The GPCP monthly analysis used in this study combines precipitation information from each source into a merged product that uses the strengths of each data type (satellite, rain gauge, etc.) and removes the biases based on hierarchical relations in the stepwise approach (Adler 2003).

The input data that is collected by GPCP includes the Global Precipitation Climatology Center (GPCC) rain gauge analyses, Microwave emission estimates over ocean, Microwave-scattering estimates over land, Geosynchronous IR-based estimates, Global Historical Climate Network (GHCN) and Climate Assessment and Monitoring System (CAMS) gauge analysis, Television and Infrared Observation Satellite (TIROS) Operational Vertical Sounder (TOVS) based estimates, and Outgoing Longwave Radiation precipitation index (OPI) (Fig. 5). The GPCC rain gauge analyses record data from 6,500 – 7,000 stations spanning from 1986 to the present. These values are interpolated using Sherman’s method (Willmott et al. 1985) to 0.50 degree resolution and then averaged to 2.5° resolution. An estimate of systematic errors in the GPCC gauge analysis (i.e. wind effects, aerodynamic effects, snow, etc.) is calculated by using bulk correction factors for monthly climatological conditions. The GHCN-CAMS gauge analysis used for January 1979 to December 1985 used the same interpolation of Sherman’s method. The same corrections for systematic error are applied. The microwave estimates over the ocean come from data recorded by the special sensor microwave/imager (SSM/I) instruments on board military satellites F8 through F15 (except for F9) and the Tropical Rainfall Measurement Mission (TRMM) satellite. The SSM/I instrument enables estimates of rain-rate. The SSM/I also
accounts for beam-filling bias due to storms and rainfall patterns typically being smaller than the satellite sensor footprint size (Wilheit et al. 1991, Chiu et al. 1993). However, the microwave-scattering estimates over land use SSM/I data at 85-GHz instead of the 19 and 22 GHz, which are used over ocean. This difference is due to the change in the background emission from ocean surfaces to continental surfaces.

The Geostationary Operational Environmental Satellite (GOES) provides infrared estimates which are used to relate cold cloud-top area to rain rate. The Television Infrared Observation Satellite Program (TIROS) supports and developed the TIROS Operational Vertical Sounder (TOVS) which is on board several NASA polar orbiting satellites. The TOVS retrieval makes a first guess for moisture retrieval using a global circulation model. This moisture is then converted to a precipitation estimate. Also, the TOVS estimates are used to fill gaps in data from the SSM/I. For the study area, TOVS data are adjusted to the bias of the zonal average mean of the SSM/I data. The OPI radiance values are directly related to higher cloud tops, which are related indirectly to increased precipitation rates. In the time period before the introduction of the SSM/I instrument (January 1979 to June 1987), OPI data is used as a replacement for SSM/I data. The OPI data that are used during this period are calibrated by the GPCP satellite-gauge estimates after the introduction of the SSM/I instrument (Fig. 5).

One of the most important aspects of the merger to create the GPCP data set is the combination of microwave and IR observed satellite data. These “microwave-adjusted IR” estimates provide a means to correct known biases arising from the indirect precipitation estimates provided by geosynchronous IR estimates. The first step of the merging method is to match the IR GPI estimates to microwave estimates to derive calibration factors, and for gaps in IR data, rainfall estimates from NOAA polar-orbiting satellites are adjusted using an
interpolation of the microwave/geosynchronous IR adjustment ratio. These adjusted coefficients are applied to the full month of GPI estimates, making the adjusted GPI (AGPI) precipitation field. Next, the average of this calculation is adjusted to agree with the large scale average of gauges where available over land. The last step is combined with inverse-error-variance weighting to produce the merged analysis (Adler 2003).

Figure 5 GPCP Merging Techniques as described in Adler 2003
Due to the difference in the instrumentation that records these inputs, the combination method is different for 1979 to 1987 and then for 1987 to the present. For 1979 to 1987, OPI estimates replace the SSM/I and geosynchronous estimates and then are merged with the rain gauge data over land. For the more recent period, 1987 to present, the combination method is designed to utilize the strengths of each input while reducing bias during each step of the merger. The TOVS estimates are used to fill in gaps of SSM/I estimates in high latitudes as they are adjusted to rain gauges as well (Fig. 5).

**CMAP**

CMAP was created in 1997 to answer the need for global, climatic precipitation data. Not surprisingly, CMAP is very similar to GPCP, whose original version was released in 1996. GPCP and CMAP both have the same temporal resolution (January 1979 to near present) and spatial resolution (2.5° degree resolution with complete global coverage). Both climatic precipitation data sets use scattering and emission microwave data from the SSM/I (after July 1987), IR data from geostationary satellites, GPCC rain gauge data, and OLR measurements in their respective merger techniques. In addition to the five source datasets that CMAP has in common with GPCP, CMAP also uses microwave sounding data from the Microwave Sounding Unit (MSU) and the enhanced version of CMAP includes the reanalysis data from the National Center for Environmental Protection and the National Center for Atmospheric Research (NCEP/NCAR) (Xie and Arkin 1997). For this study, the majority of the statistical tests will use the standard version of CMAP (the version that does not contain reanalysis data). The standard version of CMAP has more input data sources that are in common with the input data sources for
GPCP. The standard and enhanced versions of CMAP will be included in the anomaly correlation calculations. This comparison will illustrate the influence of the reanalysis data on the correlation can be measured.

The greatest difference between these datasets is in the merger and quality control methods. The merging of the source data is split into two steps (Fig. 6). The first step is to reduce random error from satellites by linearly combining the satellite estimates (GPI from GOES, microwave scatter and emission from SSM/I, OLR and OPI from geostationary satellites, and microwave data from the MSU) with the model estimates (NCEP/NCAR reanalysis) using the maximum likelihood estimation method. Over land, the random error is identified by comparing the source data to the rain gauge data over the surrounding area.

The first step does not handle the bias found in the individual sources, so the second step is to reduce the bias found in the source data. During this second step, the authors of CMAP assumed that the gauge data is unbiased (Xie and Arkin, 1997). While many consider rain gauges to have their own series of biases due to equipment or user failure, the authors of CMAP expect that these biases are minor in comparison to the biases present in the satellite data. This second step reduces the bias and merges the data together by using a method described in Reynolds (1988) where the relative distributions (i.e. the “shape”) of the blended analysis is determined by the combined analysis of the satellite source data while the amplitude is defined by the gauge analysis, where there is ample gauge data available. Reynolds (1988) outlines this process with this equation:

$$\nabla^2 \varphi = \rho.$$
Here, $\varphi$ is the in situ rain gauge data, also known as the blended field, and $\rho$ is the forcing term (or the Laplacian of the satellite sources, $\nabla^2 S$). This method eliminates the bias of the satellite data and matches the shape of satellite field to the boundary points of the in situ data.

Figure 6: CMAP Merging Procedure as described in Xie and Arkin 1997
**APHRODITE**

While GPCP and CMAP are both global climatic precipitation datasets, APHRODITE is specifically designed to report on the timing and amplitude of precipitation across the Asian continent during the monsoon and dry seasons from 1951 to 2007. Since APHRODITE seeks to mimic a climate dataset that measures precipitation during a time before satellites, APHRODITE must use the only data available: rain gauge data. The rain gauge data comes from three different sources, which include the global summary of the day (a product maintained by the National Climatic Data Center and United States Air Force), rain gauge data collected by other organizations (such as the China Meteorological Administration and the Chinese Yellow River Conservation Commission), and APHRODITE’s own collection of rain gauges in Japan and surrounding areas. Because of the emphasis that the authors of APHRODITE put on the oscillation of the monsoon season, the data is organized into daily, pentad data, and monthly data so that variations in precipitation can be seen at multiple time steps (Yatagai et al. 2012).

APHRODITE is constructed through two steps:

1. The daily climatology is created

2. The daily climatology values are used with the actual rain gauge data in the interpolation of the final data product.

The climatology is created through seven steps. First, the daily data is summed into monthly values (other than the GTS station data and only after the other stations have undergone quality control). These monthly values will provide an easier process for the second step. Second, if the monthly data and the monthly totals from the first step have over five years of data, then the monthly data and totals are averaged. This provides a way to create long term
climatology of the different rain gauge data. The third step is to prepare the climatology values (the values from the second step) at a 0.05° resolution. This is not the interpolation, but the gridding of the climatologies onto a 0.05° grid. The fourth step is to calculate the ratio between the second and third steps for each month. This ratio will be used to identify significant precipitation anomalies in the climatologies. The fifth step is to interpolate the data from step four at a 0.05° resolution using the Sherman method. The Sherman method was developed by Willmott et al. 1985, as a means to interpolate points of data on a sphere in a three dimensional space, as opposed to the two dimensional, Cartesian interpolation methods. The sixth step is to take the interpolated values from step 5 and multiply them by the mapped out averages from step three. This will multiply the ratio of interpolated anomalies with the grid of climatology data, to create a weighted value of precipitation. The final step is to take the mapped values from step six and calculate the first six components of the Fourier transform of the map, producing the daily climatology. The Fourier transformation will help to remove random or errant data.

The ratio of the daily climatology to the daily precipitation is calculated (step four) and used to weight the interpolation of the data. The ratio is defined by:

\[
\text{Ratio} = \frac{\text{daily precipitation}}{\text{daily climatology} + 1 \text{ mm per day}},
\]

and after the interpolation of the data, the data is weighted by this ratio by using:

\[
\text{Gridded daily precipitation} = (\text{daily gridded ratio}) \times (\text{climatology} + 1 \text{ mm per day}).
\]

In both instances the “1 mm per day” value was included so that values of 0 mm per day would not affect the weighting function. The weighting function had several important caveats to its calculation. First, if a ridge is between the target cell and the nearest rain gauge, the cell is given a smaller weight. Second, if a target cell is on a slope the inclines to a rain gauge, it is
given a large weight in the function. Finally, the weighting function is defined by a look up table that determines the correlation distance (Fig. 7).

Figure 7: APHRODITE merging procedure Yatagai et al. 2012
The Climate Hazard group InfraRed Precipitation with Station data archive (CHIRPS) is a quasi-global (50°S-50°N, 180°E-180°W), at a 0.05° resolution that is available from January 1981 to nearly present at pentad and monthly values. This product is generated by the U.S. Geological Survey (USGS) and University of California, Santa Barbara (UCSB) with the goal of monitoring drought and other hydrological extreme events around the world. USGS and UCSB have utilized new ground observations to build gridded precipitation climatologies at a high resolution (0.05°), which are then used to remove bias from precipitation datasets and provide a more stable approach to representing terrain-related precipitation effects. These gridded precipitation climatologies play a major role in the development of the CHIRPS product (Funk et al. 2013).

CHIRP, the former version which was first made available in 1999, is built upon three input data sources: global 0.05° precipitation climatologies, a combination of satellite-based and climate model precipitation gridded estimates varying over time, and in situ precipitation observations (Fig. 8). The climatologies are the result of merging the infrared precipitation estimates (IRP) and coupled NOAA forecast system reanalysis to create a quasi-global time series of gridded precipitation estimates. The satellite based precipitation estimates include quasi-global geostationary thermal IR satellite observations and the Tropical Rainfall Measuring Mission (TRMM) 3-hour precipitation estimate (product 3B42). The climate model used by CHIRP and CHIRPS is the NOAA Climate Forecast System v2 (CFSv2). The newest version CHIRPS, which is used in this study, has enhanced the product by including concurrent station data (Funk et al. 2013).
The UCSB Climate Hazard Group (CHG) has developed a well-documented archive of *in situ* pentad and monthly precipitation totals including observations from the monthly Global Historical Climate Network version 2 archive, the daily Global Historical Climate Network archive, the global summary of the day dataset (GSOD), and the daily Global Telecommunication System (GTS) archive provided by NOAA’s Climate Prediction Center (CPC). However, the GTS and GSOD archives have large numbers of “false zeroes” in some countries. The former version, CHIRP, was used to estimate the long-term average daily rainfall intensity in these cases. However, if the CHIRP value was higher than the long-term average daily rainfall intensity, and GTS or GSOD reported a “false zero”, the CHIRPS value was treated as a missing value.

The blending procedure used by CHIRPS comes from utilizing the older CHIRP method and the weighting of expected correlation between precipitation at a target location and the precipitation at the 5 nearest stations. This process puts a heavier weight in the algorithm on stations that are closer to the point in question, however, the authors of CHIRPS have stated that this process is “not exact and is not designed to recreate the station values precisely” (Funk et al. 2013). The weighting procedure will include values from up to four neighboring stations if the target station and the neighboring station are positively correlated. This was constructed to manage potential inaccurate observations as with the “false zeroes” recorded. This blending process is calculated at a pentad and monthly time steps and the pentad values are adjusted to match the monthly values. (Funk et al. 2013) (Fig.9).
Figure 8: CHIRP merging procedure as described by Funk et al. 2013
C. Climate Studies in the Ganges-Brahmaputra-Meghna River Basin

GPCP and CMAP have both been in use for over ten years with subsequent updates and new releases. Given their familiarity with the scientific community, these datasets both have ample amounts of documentation available, but one of the key studies was Yin et al. (2004). This study found that GPCP and CMAP strongly agree on large scale precipitation spatial patterns and temporal variations. This included important global climatic precipitation patterns like the Intertropical Convergence Zone (ITCZ) and the South Pacific Convergence Zone (SPCZ). Yin et al. (2004) also found that GPCP and CMAP agree more over land than over the ocean due to the increased presence of station data. However, Yin et al. (2004) concludes that the significant differences between GPCP and CMAP exist because of differences in the methods used to merge the source data. A very important conclusion from this study stated that caution needs to be used when viewing the data across long periods of time, as source data changes with the introduction

---

**Figure 9: CHIRPS merging procedure as described by Funk et al. 2013**
of new technology. For example, the GPI wasn’t introduced until 1986, the SSM/I wasn’t introduced until 1987, and the termination of the MSU precipitation retrievals in 1994.

Another study, Vu et al. (2011) compared GPCP, APHRODITE as well as other precipitation datasets: TRMM, PERSIANN, GHCN2, and NCEP. This study did compare the datasets to each other, but the majority of the statistics generated by this study came from comparing a SWAT hydrology model that modeled the flow in the Dakbla, Konplong, and Kontum rivers in Vietnam and used these precipitation datasets as an inputs. When APHRODITE daily precipitation values were put into the SWAT model, the generated flow was the closest to the actual observed daily flow of all datasets tested. Although there are other factors that affect the flow of a river, this finding implies that the precipitation values that are generated by APHRODITE are likely the closest to the actual daily values. On a monthly time scale, Vu et al. (2011) showed that GPCP performs well at a monthly scale at simulating flow in a river, implying that GPCP’s spatial distribution of precipitation must also be relatively accurate at a monthly scale. The findings from both of these studies were able to show that GPCP and APHRODITE provide accurate precipitation values in Southeast Asia, a region that is also greatly affected by the monsoon season, even if it is smaller and less mountainous than the GBM basin.

In another climate study by Dash et al. (2012), several precipitation and temperature datasets were compared, which included GPCP, CMAP, and APHRODITE in the Meghna basin and southern portion of the Brahmaputra basin. While this study focused more on how surface temperature affects precipitation, the individual precipitation datasets were compared to identify differences in monthly averages. This study compared precipitation datasets from 1981 to 2005, which was very close to the study period of this thesis. The conclusions drawn by Dash et al.
(2012) will have a direct relation to the results in the thesis because of this similar study period. This study showed that during the dry months (January, February, March, April, October, November, and December) from 1981 to 2005 GPCP, CMAP, and APHRODITE have very similar monthly averages. The main differences occurred during the transition months and the monsoon months (May, June, July, August, and September). During this time GPCP reported an average of up to 3 mm/day more rain than CMAP and APHRODITE. CMAP and APHRODITE had nearly the same monthly average through this study period, which is not surprising, given CMAP’s dependence on rain gauge data to determine the amplitude of precipitation.

While there are a large number of studies that use and study GPCP, CMAP, or APHRODITE, there are very few studies that have used CHIRPS because it is a relatively new dataset, first available in December 2013. In a study of seasonal drought forecasting in East Africa (S. Shukla et al. 2014) CHIRPS is shown to have a high correlation to GPCP in East Africa, although, the authors note that CHIRPS is not as strictly curated as GPCP, because the quality control of the CHIRPS data is automated (Funk et al. 2013). If this automation doesn’t detect the errors, they would be present in the published product highlighting the need to further compare CHIRPS to existing datasets to validate the data as well as check for errors.

While inter-comparisons of different precipitation data sets are not unique studies, pursuing these goals and objectives for GPCP, CMAP, APHRODITE, and CHIRPS has never been attempted in the GBM basin. These other studies though provide an opportunity to see what conclusions have already been found concerning the comparison of several of the datasets in this study. While all of the conclusions may not be applicable to this study area, the general findings will provide the framework on which this research will expand upon.
D. Needs from the scientific community

As shown by these previous studies, there are differences that exist between the datasets that need to be identified and explored for future work, such as the poor correlation during seasonal transition periods. Climate researchers need gridded datasets that include minimum and maximum temperature and precipitation, provided at daily, pentad, monthly, and annual values from years to decades. It is well understood that spatial climate patterns are most affected by terrain and bodies of water, so climate datasets need to accurately model and compensate for their effects. Even with algorithms that account for the influence of terrain and bodies of water, errors will persist. For example, the forced uplift and cooling of moist winds increases precipitation on the windward slopes and leave the leeward side of the slope dry, causing interpolation errors when methods have not been created to take topography into account. To properly capture this “rain shadow” effect, precipitation datasets need to be at a resolution of less than 100 km (Daly, 2006). This makes things difficult for global climate studies as most rain gauge station spacing is well over 100 km in many areas around the globe (Daly 2006).

The need for more spatially detailed data has been addressed with each new generation of dataset as shown with the increase in resolution from GPCP and CMAP (2.5°) to APHRODITE (0.25° and 0.5°) to CHIRPS (0.05°). As new datasets are created, the errors associated with each dataset needs to be understood as future climate studies are conducted. While this study will analyze and evaluate data at a coarse 2.5°, disagreements will still become apparent in the data.

In addition to the need for higher resolution data, there is a need for precipitation data in this region. Regional decision makers like the International Center for Integrated Mountain Development (ICIMOD, 2013), have expressed the interest in precipitation data and its effect on
the people of the GBM basin. In Akhtar et al. 2008, the authors discuss how high resolution precipitation climate data are needed for water resources in the Hindu-Karakorum-Himalaya region, which stretches from north India to the west into Pakistan. Also, Barros et al. (2006) showed the long term need for precipitation records to see the effect of rainfall on the erosion of the Himalayan region. Sarker et al. (2012) showed the need for climate precipitation data for modeling rice yield in Bangladesh. All of these studies need precipitation data as an input for impact and application models.
CHAPTER 4

METHODOLOGY

A. Data preprocessing

There are two main phases to the methodology of processing the data: a preprocessing phase of obtaining and preparing the data, and applying the tests to the data. Sections A. Data preprocessing, B. Elevation, C. Temporal Resolutions will outline the preparation of the data used. This will include a description of how the data were obtained and how Python scripts and ArcGIS were used to parse the data to the necessary temporal and spatial domain and resolution. Section D. Statistical Methods, will discuss how the individual statistical methods are calculated, how they’re interpreted and the reasoning for why each test was selected.

The four datasets were downloaded from their respective websites:

- GPCP (http://www.esrl.noaa.gov/psd/data/gridded/data.gpcp.html)
- CMAP (http://www.esrl.noaa.gov/psd/data/gridded/data.cmap.html)
APHRODITE (http://www.chikyu.ac.jp/precip/)

CHIRPS (http://chg.geog.ucsb.edu/data/).

All of the data are freely available on these websites, which is especially useful for researchers and decision makers who may not have the budget to purchase commercial data. It is very important that international products like these are shared so that anyone can pursue this research. Often agencies and organizations have restrictions or fees to obtain climatic station or gauge data, and the merged datasets provide an alternative for scientists or decision makers.

GPCP, CMAP, and APHRODITE data are available in Network Common Data Form (NetCDF), and CHIRPS is available in a GeoTIFF format. NetCDF is a set of interfaces for array oriented data access for languages like C, FORTRAN, Java, and Python. NetCDF is typically used to share and represent scientific data. NetCDF formatted data is useful for this purpose because of its unique characteristics (Rew, Unidata):

- NetCDF files contain information describing their content

- NetCDF files can be accessed by computers with different ways of storing integers, characters, and floating-point numbers

- Subsets of user-defined data can be easily accessed to get as little or as much data as needed

- Data can be appended to a NetCDF file without having to copy the dataset or redefine the file’s structure

- Multiple users can access the same NetCDF file at once
The GeoTIFF format is different from NetCDF in that this format is not stored in an array. Rather, the data is stored as a raster image that has built in projection, coordinate system datums, ellipsoids, and other spatial information encoded in the raster.

The number of files required for each dataset varied. GPCP and CMAP each only had one NetCDF file that contained an array of every month. APHRODITE, however, required a NetCDF file for each year, with an array for every day of the year in each NetCDF file. CHIRPS data required the most number of files, as a separate GeoTIFF file exists for each month of every year.

Once all of the datasets were obtained, two Python scripts (See Appendix A) were written to access the needed data from the NetCDF files and the GeoTIFF files. For the NetCDF files, two more Python scripts (See Appendix A) were written that handled the monthly gridded GPCP and CMAP NetCDF files and the daily gridded APHRODITE NetCDF. The SciPy module was used to open the monthly and daily gridded NetCDF files and write the entire, global grid for each day and month to an ESRI ASCII Grid file. However for the daily grids, a monthly average was calculated and written to an ESRI ASCII grid file. An ESRI ASCII Grid file is an ASCII file that follows a set format (see Table 2).
This file provides several lines of header information that specify the number of columns and rows in the grid, the longitude of the lower, left corner of the grid, the cell size of the grid, and the values for “no data”. Then, the data is read in an “English reading order”, starting with the top, left corner, and filling each cell in the row, going left to right. The ESRI ASCII Grid file was used for this study so that the full library of ArcGIS tools from the arcpy library could be used. This portion of the code made sure that the same temporal extent for each monthly ESRI ASCII Grid file GPCP, CMAP, and APHRODITE were used (January 1981 to December 2007) while the entire global spatial extent of both datasets were used. Since the CHIRPS data came in the GeoTIFF format, these Python scripts that convert arrays to rasters were not needed.

At this point, it is important to note that the values of precipitation are greatly increased during the monsoon season and in order to compare seasonal variations to one another there were extra lines added to the code to account for removing seasonal variations. This was accomplished by calculating the average for each month across all years of the dataset and subtracting it from individual months. For example, to remove the seasonal variability for the GPCP values of January 1981, the average of every month of January in the dataset was found and subtracted
from the GPCP January 1981 values. The result is not actual rainfall data but rather the differences of the samples from the monthly average, or the anomaly for the month of January 1981. This allows for comparison between dry and wet months without having the high precipitation values of the monsoon “wash out” the comparatively smaller variations in precipitation in the dry months.

The next portion of the code was written to use the ArcGIS Python library, *arcpy*, to utilize several data management and spatial analytical tools. A polygon shapefile was created using the “Make Fishnet” tool to match the extent of the GBM basin at a 2.5° resolution. This grid originally contained 55 individual cells which comprised of 5 rows and 11 columns. As seen in the next section, several cells were removed from these 55 cells to better represent the GBM basin, so a final polygon shapefile with 38 cells was used for analysis. The “Clip” tool was used to select only the raster data that was spatially located in this grid. This grid spatially matches up exactly with the cells of data in the GPCP, CMAP, APHRODITE, and CHIRPS datasets so that when the “Clip” tool was used, there were no overhanging or overlapping raster cells in any of the grid cells. This ensured that the data was in the same extent of the grid. The “Clip” tool was used again for the CHIRPS data, and while the edges match up exactly with rectangular extent of the grid, CHIRPS has a much higher density of data (2,500 cells of raster data per grid cell) compared to GPCP, CMAP (1 cell of raster data per grid cell), and APHRODITE (100 cells of raster data per grid).

To compensate for this resolution, an average of 2,500 CHIRPS cells and the 100 cells of APHRODITE raster data was taken that match the spatial extent of a single GPCP cell. The “Zonal Statistics as a Table” tool was used to calculate the average of all CHIRPS cells within each cell of the grid. Several other descriptive statistics were recorded, such as the standard
deviation of CHIRPS values within the grid cell, the minimum and maximum CHIRPS values within the grid cell, and the median CHIRPS value in the grid cell. These tables were added to the same geodatabase as the GPCP data.

B. Elevation

The extreme range of topography of the GBM basin is a major factor in the climatology and monsoon season of this region. For this study, an accurate Digital Elevation Model (DEM) was utilized so that the effects of topography on precipitation could be mapped and studied. There are numerous DEMs available, and the Hydrological data and maps based on Shuttle Elevation Derivatives at multiple Scales (HydroSHEDS) was used in this study because it is a DEM that is designed for hydrologic applications. HydroSHEDS is a dataset that takes elevation data from the Shuttle Radar Topography Mission (SRTM) and was improved upon by using algorithms that included filling voids, flow direction, and hydrologically conditioning the rasters. When hydrologically conditioning the DEM rasters, HydroSHEDS forced the DEM to produce the correct river network topography while preserving as much of the original SRTM information as possible. The HydroSHEDS data is globally available at 3 arc-second (approximately 90 meters) to 5 minute resolution (10 kilometers) (Lehner et al. 2006). For this study, the hydrologically conditioned, 3 second DEM was used (Fig. 10).
Figure 10: HydroSHEDS DEM of the GBM basin overlaid with the study grid, the outline of the basin, and the location of the three rivers (Lehner, 2006)

In order to use the HydroSHEDS, the appropriate rasters needed to be downloaded from the HydroSHEDS website (http://hydrosheds.cr.usgs.gov/index.php). The global DEM data are split into 5° x 5° tiles, so each tile had to be downloaded from the website. Then, all of the tiles were opened into ArcMap and using the “Mosaic” tool, the tiles were joined together. Now that a complete single DEM raster of the study area was available, the “Zonal Statistics as a Table” tool was used along with 38 cell polygon to calculate the average elevation within each cell. Based on this average elevation value, a simply classification was conducted so that each cell was classified as “Low Lands”, “Transitional Lands”, and “Mountainous Lands”. This allows the statistical tests performed for each cell, to differentiate between elevations (Fig. 10).
C. Temporal

Each of the four datasets has a unique temporal availability. For GPCP and CMAP, there are monthly values available from January 1979 to near present; CHIRPS offers monthly values from January 1981 to near present; and APHRODITE offers daily values from January 1, 1951 to December 31, 2007. Since the goal of this study is to compare the data of each merged dataset, the study period had to be selected during a period that all four datasets share. Because of this requirement, the months from January 1981 to December 2007 were used for this study.
To address another goal of the study, the monsoon and dry season needed to be defined. For the time period between 1981 and 2007, the monthly averages were calculated to observe the influence of monsoon season. Barros and Lang (2002), noted that the start of the monsoon season in Nepal begins in early June, which was seen in the monthly averages for the entire GBM basin, and from the data it appears that the monsoon season lasts till September. Therefore, for this study the months of June, July, August, and September were defined as being the monsoon season, while the remaining eight months, January, February, March, April, May, October, November, and December, are defined as the dry season. There is rain present during the dry season. In comparison to the monsoon season, it is significantly less.

D. Station Data

In addition to elevation and seasonal differences, another key factor that could result in differences in data, is the varied number of stations used in the merged datasets. This difference could greatly affect the agreement or disagreement between the datasets, especially CMAP and APHRODITE, who’s merging process relies heavily upon the amplitude of rain gauge. GPCP and CMAP both use GPCC rain gauge station data, APHRODITE uses rain gauge data from local regional governments and meteorological groups, and CHIRPS data uses national and regional meteorological stations. The maps show the location of the average number of stations used over the time period (Fig. 12, 13, and 14). All rain gauge locations have been commonly binned to 2.5° resolution. From these maps, it becomes obvious that there are clear differences in the density of the various rain gauge networks. Through all three datasets, there is a dense network of rain gauge data available through Nepal. Also it is noteworthy to mention that APHRODITE and CHIRPS have a high density of rain gauges in Northern India. All three of the rain gauge networks have a lowered density of stations in the northern grid cells of the study.
region which overlay the Himalayas. This lack of rain gauge density in the mountainous regions could lead to lowered correlations between datasets in these higher elevations.

Figure 12: Average APHRODITE station density over the entire study period.
Figure 12: CHIRPS average station density for the entire study period.

Figure 13: GPCC average station density for the entire study period.
E. Statistical Methods

*Correlation*

One of the most widely used statistical values used to measure association between two variables is Pearson’s Correlation. Correlation can be thought of as the ratio of the sample covariance of the two variables to the product of the two standard deviations (Wilks, 2011). Correlation is not as robust or resistant as the other tests used in this study; a correlation coefficient doesn’t necessarily recognize nonlinear relationships between variables and correlation coefficients can be sensitive to one or a few outlying point pairs (Wilks, 2011). However, despite this weakness, correlation is commonly used because its form is easily manipulated mathematically and correlation is closely related to regression analysis, the bivariate and the multivariate Gaussian distributions.

There are two major characteristics of correlation. First, the correlation coefficient is bounded by -1 and 1. These bounding values describe the linear association between two points. A perfect, linear relationship would have a value 1, meaning that the values from dataset x will always increase in relation to an increase of the values from dataset y. Likewise, a correlation coefficient value of 1 would also mean that as the values from dataset x decrease, the values from dataset y would also be decreasing. Conversely, a perfect negative linear relationship would have a value of -1, meaning that the values from dataset x will always increase in relation to a decrease of the values from dataset y or if the values of dataset x were to decrease, and then the values of dataset y will increase. Secondly, correlation is often represented by r, but can also be represented by $r^2$. This squared value is often described as a factor to quantitatively describe how strongly one variable “explains” the other variable. For this study, correlation was used to
determine the monsoon and dry months and to measure agreement and disagreements as time progresses.

**Empirical Distribution Function (EDF)**

This graphical representation uses the vertical axis as the cumulative probability estimate associated with data on the horizontal axis. In other words, the plot represents relative frequency estimates for the probability that a random future event will not exceed the value on the horizontal axis. For example, with precipitation anomalies along the x-axis and the cumulative probability on the y-axis, the value on the y-axis determines the probability that a random event will not exceed a certain amount of precipitation. An S-shaped Empirical Distribution Function is representative of a reasonably symmetric distribution, with comparable numbers of observations of data on either side of the median. The Empirical Distribution Function can be represented by:

$$p(x_{(i)}) = \frac{i - a}{n + 1 - 2a}, 0 \leq a \leq 1$$

Here in this equation $p(x)$ is the empirical distribution function, and $i$ is used to rank the order of the statistics $x_{(i)}$. As with many statistical equations, $n$ is the sample size. Different values for $a$ result in different plotting position estimators. There are numerous different estimators, such as $a = 0$ which is used to approximate the mean of sampling distribution, $a = 0.3$ or $a = 1/3$ to approximate the median of sampling distribution, or $a = 1$ to approximate mode of sampling distribution. These estimators are typically used to describe the characteristics of the sampling distributions of the cumulative probabilities associated with the order statistics. For this study, $a = 1/2$ was used to approximate midpoints of n equal intervals on $[0, 1]$ in order for the
plotting position formula to use the sampling distribution of data quantiles. $x_i$ which corresponds to particular, fixed cumulative probabilities (Wilks, 2011).

**Rank Correlation Coefficient**

The Spearman Rank Correlation Coefficient allows for nonparametric data to be evaluated for linear association or for correlation between two independent variables. As a nonparametric technique, this method is unaffected by the distribution of the population. Another strength of the Spearman Rank Correlation Coefficient is that small sample sizes can be used, and the method is straightforward to apply. It is similar to the Pearson’s Correlation Coefficient. While Pearson’s Correlation Coefficient using the raw data, the Spearman Rank Correlation Coefficient uses ranked data. Note that, if the data are normally distributed, the Pearson Correlation Coefficient is more powerful than the Spearman Rank Correlation Coefficient.

The Spearman Rank Correlation Coefficient is calculated by ranking each variable of both data sets from lowest to highest and the difference between each pair of data is calculated. If the sum of the square of the difference of the ranks is small, the datasets are highly correlated; in other words, the magnitude of the sum of the square of the difference between ranks is related to the significance of the correlation. The Spearman Rank Correlation method can be explained by this equation:

$$
\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}
$$

Where $d_i^2$ is the difference between ranks, and $n$ is the number pairs of variables.
The significance of a Spearman Rank correlation can be calculated but it must be assumed that there is no correlation between the two variables and comparing the correlation value to a table of critical values. For these data, a t-statistic, a measure of distance of a value from its standard deviation, was used and compared to a table. The t-statistic was calculated with the following equation:

\[ t = \frac{r_s \sqrt{n - 2}}{\sqrt{1 - r_s^2}} \]

Where \( r_s \) is the correlation and \( n \) is the number of pairs of variables.

When used for detecting trends in a time series, seasonal variation is one of the major issues with the Spearman rank correlation coefficient, as seasonal variations in a data set can hinder the detection of long term trends. To overcome this, the seasonal variations were removed before applying the test. The seasonal variations were removed by finding the average value of the month over the study area and 30 year period, and then this average was subtracted from each individual month’s value of precipitation.

The interpretation of the Spearman Correlation relies on several factors. For instance, the sign of the Spearman Correlation Coefficient indicates the direction of association between the independent and dependent variable, or in other words the ranking of the values of dataset x and the rankings of dataset y. If the dependent variable increases as the independent variable increases, the Spearman correlation coefficient will be positive. If the value of the Spearman correlation coefficient is one, then the dependent and the independent variables are perfectly monotonically related, meaning that the derivative of the two pairs of data will always have the same sign. However, when the trends (increasing or decreasing) of independent and dependent
variable are opposite to one another the Spearman correlation coefficient will be negative. If the dependent and the independent variable have no tendency for increasing or decreasing the Spearman Correlation Coefficient will be zero (Hollander, 1999).

Anomaly Correlation:

Although often used to measure the quality of a forecast system, the anomaly correlation is a useful statistical measure in this study. Anomaly correlation correlates the anomalies (the difference between the climate average and the actual value of the data set) rather than the actual values. The anomaly correlation is sometimes referred to as a pattern correlation because it was created to identify similarities in the patterns of anomalies of datasets (Pearson, 2013). Also, because of the way the anomaly correlation is calculated, it does not penalize conditional or unconditional biases, which means that anomaly correlation is not a measure of actual skill, but rather the potential skill of one dataset to agree with another dataset. (Wilks 2011)

Anomaly correlation is typically used to measure how well a forecast matches an observed result. Because this study assumes that none of the four datasets are a “true observation” (rather this study assumes that each dataset is a merged dataset constructed with specific assumptions that their original authors hold to be most important in the climate datasets creation), the anomaly correlation is used to measure the agreement between the anomalies of the climate datasets at annual time steps. Unlike the Spearman’s Rank Correlation, which was calculated over the entire time period, the anomaly correlation was used to measure the agreement between climate precipitation datasets from year to year. This was done to see if agreement increased with time as instrumentation improved.
The first step in calculating the anomaly correlation is to calculate the anomalies, which is done by subtracting the climate average from the actual values of the two datasets. This is done for every spatial, grid cell of data for each dataset. These values are multiplied and averaged, then divided by the square root of the product of the average squared differences, as shown in the equation below where \(f\) is the forecasted value, \(c\) is the climate average, and \(a\) is the actual value (Persson, 2013):

\[
ACC = \frac{(f - c)(a - c)}{\sqrt{(f - c)^2 (a - c)^2}}
\]

In this study, the \(f\) and \(a\) are going to be two arbitrary datasets, so that their precipitation patterns can be compared to one another. The variable \(c\) is the monthly average over the entire dataset. Essentially, the value within each set of parenthesis is the anomaly value of each dataset and the bar denotes that the value is averaged over the domain.

When comparing datasets, an anomaly correlation value of 0.6 is generally considered to be the lowest threshold of determining the effectiveness of one dataset being used to forecast, or predict, another dataset. This value does not come from a specific calculated value but rather a consensus of experienced opinion (Krishnamurti, et al 2002). As the results of this calculation are discussed, keeping this value in mind will be important in determining how datasets spatially agree or disagree.

All of the statistical methods discussed in the section were calculated using Microsoft Excel. Once the Python scripts organized the data, and found the values for the data in each cell,
a Python script was written to convert the cell values into a text file, and finally the data was
opened into Excel. Here, tables were created to generate the intermediate steps for each method,
so that a final table of the results from these tests could be produced. Excel was used to process
these statistical tests because of the small number of cells in the area, 38, and the ease of data
manipulation in Excel. If this study were to include a larger area, or a higher resolution grid, then
the data would have been left in arrays and calculated in Python.
CHAPTER 5

RESULTS

A. Introduction

The findings have been organized by each statistical method. This was done to show the significant characteristics of each dataset from broad calculations that take into consideration the entire area of the dataset over the entire 27 year study period to individual elevations during certain seasons of specific years. First, the Empirical Distribution function was used to address Objective 1; find a broad measure that describes the general behavior of each dataset over the entire area for the entire 27 years of the study period. Objective 1 was further explored by using Pearson’s correlation to examine how individual datasets compare to one another over the entire study area and period. Then, Objective 2, which was to find a method of correlation that will describe agreement and disagreement at different elevation zones, was addressed by using the Spearman’s Rank Correlation to show how the agreement between datasets changes depending on spatial differences, as defined by elevation categories. Objective 3, which was to find a
correlation that will describe agreement and disagreement during the dry and monsoon season, was accomplished by using the Spearman’s Rank Correlation and the Anomaly Correlation to highlight the unique differences between the datasets in these two seasons. Following this, Objective 4, which was to determine if the density of rain gauges in the development of datasets have affected the correlation between the datasets, was addressed by graphing the monthly anomaly correlation between two datasets at a monthly time step against the number of rain gauge stations used by each dataset. This was done so instances could be identified where a lack of rain gauge data seemed to affect the anomaly correlation. These graphs were also plotted so that the four months (June, July, August, September) of the monsoon season and the eight months (January, February, March, April, May, October, November, December) were plotted separately at each of the three elevation classes. Finally, the anomaly correlation of the entire wet and dry seasons for each year were plotted over the 27 year study period to highlight the changes in agreement between datasets over time. This also helps to address the overarching goal of the study, to generate a measure of correlation among climatic precipitation datasets, as well as address the specific need of Objective 2 to highlight how agreement between individual datasets changes between monsoon and dry seasons at differing elevations.

B. Empirical Distribution Function

The Empirical Distribution Function (EDF) is used to show the probability of a random precipitation event to exceed a specific rain rate, where the specified rain rate (or anomaly) is plotted against the x-axis and the probability of non-exceedance is plotted on the y-axis. The EDF is calculated for the entire study area and period. In the graphs generated below (Figs. 14, 15), the x-axis is the anomalies, or the distance from the average rainfall event. In other words, a
value of 0 represents an average rainfall event, while a value of -1 represents a rainfall event that is one millimeter per day less than the average rainfall.

To illustrate how the Empirical Distribution Function is interpreted, notice how CHIRPS has a lower probability of non-exceedance than GPCP, CMAP, and APHRODITE for random rainfall events from the average (0) to an anomaly of negative one millimeter per day (-1 mm/day). Because of this, CHIRPS has a lower probability of reporting a negative anomaly between 0 and -1 mm/day, for the entire study area and period, than the other datasets. In other words, CHIRPS is reporting that negative rainfall anomalies of 0 to -1 mm/day are less likely to occur than for the other three datasets. Furthermore, CHIRPS and APHRODITE anomalies near -1 mm/day have very similar non-exceedance probabilities, which are both less than GPCP and CMAP, whose non-exceedance probabilities are nearly equal as well.

Another approach for interpreting this graph is to consider that the y axis is the normalized sum of the frequency of the occurrence of the values on the x-axis. In other words, the y axis represents the percentage of the total number of values of the dataset that lie on the curve to the left of the point on the x axis. For example, if a dataset’s EDF had a 20% probability for a -1 mm/day anomaly, then 20% of all values from the dataset are of a -1 mm/day anomaly or lower. Further, given that the values for anomalies are plotted along the x-axis, the zero value is the average value of each datasets and is illustrated on the x-axis by the division between the positive and negative values on the x-axis to represent the positive and negative anomalies in each dataset. Figures 15 and 16 have been separated accordingly to highlight how each dataset handles negative and positive anomalies, respectively.
Figure 15: Negative Anomalies for the Empirical Distribution Function

From Fig. 15, there are several differences that are apparent. First, from 0 to -1 mm/day, CHIRPS has a lower sum of frequency than the other datasets. However, at -1 mm/day APHRODITE and CHIRPS have the same sum of frequency of 0.11, while GPCP and CMAP share the same sum of frequency of 0.14. These calculations can be interpreted as 11% of all rain events for APHRODITE and CHIRPS had a -1 or lower mm/day anomaly, while 14% of all rain events for GPCP and CMAP for these same anomalies, meaning GPCP and CMAP are reporting more negative anomalies that are less -1 mm/day than APHRODITE and CHIRPS. Interestingly, beyond -1 mm/day, APHRODITE is reporting lower sum of frequencies (0.02 at -3 mm/day) than the other three datasets (0.03 at -3 mm/day), until they converge at approximately -6 mm/day. These lower probabilities mean that APHRODITE is reporting fewer of these more extreme negative anomalies, than GPCP, CMAP, or CHIRPS.
There is less variation in the positive anomalies, which is probably due to all four of the datasets reporting the heavy rainfall of the monsoon (Fig. 16). However, it is interesting to note that CHIRPS’ and APHRODITE’s non-exceedance probabilities are nearly equal and GPCP and CMAP non-exceedance probabilities are nearly equal as well, thus it can be concluded that over the entire time period and study area, the pairs of CHIRPS-APHRODITE and GPCP-CMAP behave very similarly. This likely comes from the two pairs of data having similar spatial resolution (2.5° for GPCP-CMAP and 0.05° and 0.25° for CHIRPS-APHRODITE, respectively).

From the negative and positive EDF graphs, it can be observed that CHIRPS’ and APHRODITE’s curves are more vertical near the average (a range from -1 mm/day to +1 mm/day), than GPCP and CMAP. This vertical nature of the graphs show that CHIRPS and APHRODITE are reporting more anomalies that are closer to the climatic average than GPCP and CMAP. In other words, rainfall events recorded by CHIRPS and APHRODITE have a higher probability of having an anomaly very close to the climatic average than GPCP and CMAP. For example, CHIRPS and APHRODITE have a sum of frequency of 0.94, meaning that 94% of all records for CHIRPS and APHRODITE are 2 mm/day or lower. Meanwhile, GPCP and CMAP have a sum of frequency of 0.93, meaning 93% of all records for GPCP and CMAP are reporting 2 mm/day or less.
It is interesting to note that for all negative and positive anomalies, GPCP and CMAP have nearly identical non-exceedance probabilities. This shows that when observed over a broad area and time period, CMAP and GPCP report similar probabilities of rainfall anomalies. This is likely due to GPCP and CMAP sharing many of the same source input data, such as SSM/I microwave emission and scattering estimates, GPI and OPI IR indices, and GPCC rain gauge analyses.

As previously stated, APHRODITE and CHIRPS tend to report more anomalies that are closer to average than GPCP and CMAP. It should be noted that higher resolution data of APHRODITE and CHIRPS were averaged to the resolution of GPCP and CMAP before this calculation was made. Therefore, the average of APHRODITE and CHIRPS was the value tested in the EDF and not the actual values of APHRODITE and CHIRPS. This averaging was done so that the comparisons could maintain equal spatial resolutions. This should be considered
C. Correlation

The next step to analyzing the data is to study how each data set relates to one another spatially. To accomplish this, a correlation coefficient for each cell of the study area was calculated and mapped in order to spatially analyze where datasets agreed and disagreed over the study area (Fig. 17). This correlation is for the entire period from 1981 to 2007.

Figure 17: Correlation Maps from the entire study period
Several conclusions can be drawn from these maps from this pair-wise comparison. First, the three maps in the left column correlate CMAP with one of the other datasets. These maps show that CMAP lowers the correlation of the pair-wise comparison with any of the datasets in this study. While the correlations between CMAP and the other datasets in the low lands over central India are in agreement (correlation values of 0.8 or greater), but in transitional and mountainous elevations the correlation between CMAP and the other datasets are significantly lower.

Meanwhile, GPCP and APHRODITE have the highest correlation among any of the datasets over the entire study area. The correlations of GPCP-CHIRPS and CHIRPS-APHRODITE are both in good agreement (correlation of 0.5 or higher). These maps show that over the entire study period and over varying topographies GPCP, CHIRPS, and APHRODITE have higher correlations to their respective parings than to the parings that contain CMAP.

In addition to the spatial patterns of temporal correlation, basin-area averages of temporal correlations were calculated (Fig. 18) and graphed to show the change in correlation over time. This technique shows the annual correlation of each dataset pair for its annual average of all seasonally removed monthly values of precipitation. There are several conclusions to be drawn from this graph as well.
First, the correlations involving CMAP are the lowest among the six different pair-wise comparisons. The correlations with CMAP are at their lowest in 1984, nearing zero and even reaching slightly negative values. Meanwhile, the other correlations of each pair of datasets were above values of 0.8. This means that the recorded rainfall records of CMAP are almost completely independent of the recorded rainfall of the other datasets during this year. This is further seen in the graph of the annual anomaly graph. Here, CMAP is seen having a negative anomaly while the other three datasets have a positive anomaly. Finally, in 1997 there is a clear drop in correlation for nearly all of the comparisons, except for the GPCP-APHRODITE comparison. This implies that there was a discrepancy in the merging techniques of the datasets, which could possibly linked to a lack of stations, a situation which will be further described in the next section. While APHRODITE’s input data come entirely from rain gauges, it has such a robust collection of rain gauges that even during a period of reduced stations; APHRODITE still has an average of at least 30 stations per grid in the study area. The merging technique of GPCP
puts a stronger emphasis on satellite data for the amplitude of rainfall than that of CMAP and CHIRPS. Because of the high correlation between GPCP and APHRODITE, GPCP’s collection of satellite input source data must match well to the dense rain gauge network of APHRODITE during this year.

The last graph of Pearson’s Correlation (Fig. 19) shows correlation of area averaged quantities where the average monthly climatological precipitation has first been removed. The main conclusion to be made from this graph is that the monsoon season has a clear effect on the correlation between datasets. During the drier months (January, February, March, April, May, October, November, December) the correlations have a relatively higher correlation value, with little variation, while the correlation values of the monsoon months (June, July, August, September) are very divergent and typically have a much lower correlation than the monsoon season. It can be argued that the monsoon begins in May based on the sharp decline seen by the CMAP related correlations. However, given the errors prone to CMAP from the other correlations shown in this section, the early sharp decline in CMAP in the month of May is ignored since it is not as prevalent in the other datasets.
D. Spearman Rank Correlation

The Spearman Rank correlation was calculated to address spatial differences in the correlation of the datasets, as represented by the elevation categories. Since the Spearman Rank is a nonparametric statistical method, the values calculated will give a stronger value of agreement or disagreement because of the non-parametric nature of precipitation (Wilks, 2011). Further, for this statistical test the data was separated into Monsoon and Dry seasons (Table 3 and Table 4, respectively) and the various elevation classifications. These values represent the entirety of both seasons and the entire class of each elevation. For example, the Spearman’s

Figure 19: Seasonal Correlations between each dataset
Rank Correlation value for GPCP-CMAP for the Low Lands during the Monsoon season represents all 17 cells in the Low Lands during only June, July, August, and September.

One of the main conclusions to be drawn from these tables is that all of the datasets have a high level of agreement in the lower elevations (Spearman’s Rank Correlation value of at least 0.7), but during both the Monsoon and Dry seasons, as the elevation increases the Spearman’s Rank Correlation decreases. Another main conclusion from these tables, is that in almost every elevation and dataset pairing, the Dry season out performs the Monsoon season with higher correlations. This would suggest that the datasets are in better agreement for determining the lack of rain rather than the presence of rain, which will be important for determining periods of drought. Finally, one of the most important conclusions is that during the monsoon season APHRODITE has a Spearman’s Rank Correlation near 0 when related to the other datasets in mountainous regions. This would suggest that APHRODITE’s data is independent of the other datasets, thus APHRODITE is reporting significantly different values of monsoon rainfall in the Himalayas than the other datasets.
<table>
<thead>
<tr>
<th>Monsoon Pairings</th>
<th>Low Lands</th>
<th>Transition Elevations</th>
<th>High Lands</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPCP – CMAP</td>
<td>0.7776</td>
<td>0.4387</td>
<td>0.4891</td>
</tr>
<tr>
<td>GPCP – CHIRPS</td>
<td>0.7898</td>
<td>0.6767</td>
<td>0.6410</td>
</tr>
<tr>
<td>GPCP – APHRODITE</td>
<td>0.9129</td>
<td>0.8459</td>
<td>0.0318</td>
</tr>
<tr>
<td>CMAP – CHIRPS</td>
<td>0.7471</td>
<td>0.4303</td>
<td>0.5894</td>
</tr>
<tr>
<td>CMAP – APHRODITE</td>
<td>0.7651</td>
<td>0.4464</td>
<td>-0.0210</td>
</tr>
<tr>
<td>CHIRPS – APHRODITE</td>
<td>0.7860</td>
<td>0.6516</td>
<td>0.0214</td>
</tr>
</tbody>
</table>

Table 3: Spearman’s Rank Correlation Values during the Monsoon Season

<table>
<thead>
<tr>
<th>Dry Pairings</th>
<th>Low Lands</th>
<th>Transition Elevations</th>
<th>High Lands</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPCP – CMAP</td>
<td>0.8519</td>
<td>0.7151</td>
<td>0.6308</td>
</tr>
<tr>
<td>GPCP – CHIRPS</td>
<td>0.8022</td>
<td>0.7539</td>
<td>0.6290</td>
</tr>
<tr>
<td>GPCP – APHRODITE</td>
<td>0.9337</td>
<td>0.9029</td>
<td>0.7908</td>
</tr>
<tr>
<td>CMAP – CHIRPS</td>
<td>0.7656</td>
<td>0.7174</td>
<td>0.6005</td>
</tr>
<tr>
<td>CMAP – APHRODITE</td>
<td>0.8401</td>
<td>0.7376</td>
<td>0.6883</td>
</tr>
<tr>
<td>CHIRPS – APHRODITE</td>
<td>0.8001</td>
<td>0.7885</td>
<td>0.7020</td>
</tr>
</tbody>
</table>

Table 4: Spearman’s Rank Correlation Values during the Dry Season
E. Stations and Anomaly Correlation

While Spearman Rank’s provides nonparametric analysis of data, Anomaly Correlation is a statistical test that determines the correlation of the spatial patterns of precipitation. By overlaying the number of rain gauges from each dataset and the Anomaly Correlation, the effect of the presence of rain gauge data can be observed. There are a total of 36 graphs that illustrate the results and are found in Appendix B. For brevity, the most important conclusion will be discussed here with only one or two noteworthy examples for each.

It is important to note that the spatial distribution of the rain gauges widely varies over time from dataset to dataset. The average annual number of rain gauges was graphed and included in Fig. 12, 13, and 14 to illustrate their distribution in relation to elevation. Also, the enhanced version of CMAP, which includes reanalysis data, was included in these tests. This allows the standard and enhanced CMAP to be compared against one another based on the correlations to the other datasets. This assessment will show that the pair-wise comparisons including the enhanced CMAP data have correlations much closer to zero than the pair-wise comparisons including the standard CMAP data.

First, GPCP and the standard CMAP datasets both share the same GPCC rain gauge data. However, due to the difference in GPCP and CMAP’s merging techniques, the effect of the presence of rain gauges on the Anomaly Correlation is very different. For instance, during the dry season in the low lands, CMAP-APHRODITE has a clear decrease in Anomaly Correlation from 1980 to 1985 when there are very few GPCC rain gauges on average (Fig. 21). However, for the equivalent season and elevation the GPCP-APHRODI E pair-wise comparison does not experience this level of lowered or highly varied Anomaly Correlation (Fig. 20).
Even though the presence of rain gauges does have some effect on how well the standard CMAP dataset agrees with APHRODITE, there are other instances where the correlation values significantly change but there is either little change in station density (see the drastic lowering of correlation in Fig. 21 from 1990 to 1991) or there is a drastic increase in station density (see
August 1998 in Fig. 21 when despite the increase in APHRODITE stations, there is a sharp decrease in correlation. It is not explicitly obvious as to why the correlation changed during these first instances given that the rain gauge density didn’t significantly change. This disagreement likely comes from the satellite based input source data disagreeing with the data from the APHRODITE stations. In the second instance, the sudden increase in APHRODITE rain gauge station density would have been more sensitive to changes in rainfall rates due to localized, convective storms, which a less dense CHIRPS rain gauge network would miss. Overall, these graphs show that the presence of GPCC rain gauges can affect the correlation of GPCP and CMAP (standard), but they are not the only input source that could be causing the correlation to change. These statistics represent area-averaged behavior so it is likely that important rainfall events don’t occur evenly over the domain and thus, are sampled differently by the different gauge networks.

The enhanced CMAP dataset was correlated with the other datasets, and there are several main points to be drawn from these comparisons. First, the enhanced CMAP dataset was correlated to GPCP, to see how the reanalysis data has changed the correlation values. During the monsoon season in the lower elevations, the correlation between CMAP-Enhanced and GPCP (Fig. 22) is much more varied than CMAP-standard and GPCP (Fig. 23). Fig. 22 shows that density of rain gauges do not seem to have an effect on correlation to GPCP. However, Fig. 23 shows that correlation is lower during a period of less dense rain gauges, in a period before microwave scattering and emission satellite data was available. The availability of microwave data is important because even though there is a decrease in the average density of rain gauge stations from 1997 to 2001, the correlation remains high (over 0.6) during this time, because microwave data is available in addition to infrared data.
Figure 22: CMAP-Enhanced correlated with GPCP, with GPCC rain gauge density for lower elevations in the monsoon season

Figure 23: CMAP-standard correlated with GPCP, with GPCC rain gauge density for lower elevations during the monsoon season
During the dry season, there are similar results between the comparisons in correlation values between CMAP-Enhanced and CMAP-standard. Fig. 24 shows that the density of rain gauge data does not have an effect on correlation between CMAP-Enhanced and GPCP. Similar to Fig. 23, Fig. 25 shows that there are lowered correlations from 1981 to 1986, but the correlations improve when the rain gauge station density is increased.

**Figure 24:** CMAP-enhanced correlated with GPCP, with GPCC rain gauge density for transition elevations during the dry season

**Figure 25:** CMAP-standard correlated with GPCP, with GPCC rain gauge density for transition elevations during the dry season
Further, the correlation between the enhanced version of CMAP and another dataset can be affected by the absence of rain gauge data from the other dataset. In both the transitional (Fig. 26) and the mountainous elevations (Fig. 27), the CHIRPS density of rain gauge data drops to zero during August of 1999. In both of these instances, the correlation drops to nearly -1, indicating that the infrared satellite products used by CHIRPS recorded an anomaly of opposite sign from the merged product of the enhanced CMAP dataset. It is interesting to note, that during August of 1991 the density of CHIRPS rain gauge data drops to zero as well, but the correlation does not drop to -1 in either case (Fig. 26 and Fig. 27).

![Figure 26: CMAP-standard correlated with GPCP, with GPCC rain gauge density for lower elevations during the monsoon season](image)
The last comparison for the enhanced CMAP product is to APHRODITE. Because APHRODITE relies entirely on rain gauge data, the authors of APHRODITE have created the largest network of rain gauge data. During the entire study period the density of rain gauge data from either dataset has positively affected the correlation (Fig 28). Even during the four year period from 1997 to 2001, where there could be more than 200 stations per grid cell, the correlation actually decreased to negative values. These negative values occurred because the high APHRODITE rain gauge density recorded an anomaly with a different sign from the merged enhanced CMAP product. In summary, the addition of the NCEP/NCAR reanalysis data considerably degrades the skill of CMAP. This may be due to the fact that the reanalysis cannot adequately capture rain events at the 2.5 degree spatial scale.

**Figure 27: CMAP-enhanced correlated with CHIRPS, with GPCC and CHIRPS rain gauge density for mountainous elevations during the monsoon season**
CHIRPS has varying effects on the Anomaly Correlation depending on the dataset that it is paired with. For example, when paired with standard CMAP, the slow increase in CHIRPS rain gauge density in the low lands during the dry season from 1997 to 2007 is reflected in a slow increase in correlation (Fig. 29). However, there are instances when the influence of the CHIRPS rain gauge network isn’t as strong. This is seen in the CMAP-CHIRPS comparison in mountainous regions in the dry season (Fig. 30). The sharp decrease in CHIRPS station density from 1995 to 1997 does not seem to effect the correlation, because the correlation increases during this period. Since the density of CHIRPS rain gauge data has a more positive effect on correlation in the lower elevations than in mountainous regions, it can be concluded that the CHIRPS weighting procedure of incorporating rain gauge data does not account for elevation changes or account for the low density of rain gauge data in mountainous regions.
Figure 29: Low Lands, Dry Season CMAP-CHIRPS Anomaly Correlation with Average Rain

Gauges per cell

Figure 30: Mountainous Land, Dry Season CMAP-CHIRPS Anomaly Correlation with Average Rain Gauges per grid cell
APHRODITE’s robust collection of rain gauges does have an effect on the anomaly correlation. The density of APHRODITE rain gauges has an effect on correlation that can be seen when compared to the number of stations from GPCC (Fig. 31). When there are a large number of APHRODITE stations available in mountainous regions then APHRODITE can increase the Anomaly Correlation more so than its other paired datasets. This can be seen in the comparison between GPCP-APHRODITE during the dry season in the mountainous region. In this graph, the peak of APHRODITE stations coincides with the increase in Anomaly Correlation (Fig. 31). However, in lower elevations, correlations involving APHRODITE are not as affected by the density of APHRODITE rain gauges. For example, in Fig. 23 APHRODITE is being compared to CMAP in the lower elevations during the dry season. This graph shows that the correlation is lowered from 1981 to 1986 when there is a lack of GPCC rain gauge data, but during this time there are on average approximately 40 APHRODITE rain gauges. This drop in correlation is caused by the lack of GPCC rain gauge data. Further, the there was a drop in correlation in 1998, despite there being over 200 APHRODITE stations. But again there was a drop in GPCC rain gauge density. It can be concluded that in the low lands, the correlations are more influenced by the lack of other rain gauges than abundance of APHRODITE rain gauges.
To further compare the datasets, the entire monsoon and dry seasons for each year are compared to one another at each elevation. This was done so that general trends between the correlations could be identified. First, in Fig. 32 the dry seasons in the lower elevations were compared. In 1984 and 1990, there is a clear decrease in correlation with all pair-wise comparisons that include the standard CMAP. As time increases through the study, the correlations involving the standard CMAP dataset improve (see the correlation values from 1999 to 2007 in Fig. 32). However, the enhanced CMAP has very low correlation values throughout the study period when compared to GPCP, CHIRPS, and APHRODITE. These near zero values indicate that the enhanced CMAP is not in agreement with the other datasets because of the addition of the reanalysis data.
At the next elevation (transitional elevations, Fig. 33), the anomaly correlation shows very similar results to the low land correlation values, except the standard CMAP correlations are at lower values. For example, from 1999 to 2007 in the dry season, over the lower elevations, the correlation ranges from 0.8 to near 1.0. However, during this same time period over the transitional elevations (Fig. 33), the correlation ranges from 0.6 to near 1.0. Interestingly, the correlations involving enhanced CMAP are still low, but the values are much more varied, ranging from -0.6 to 0.2. Despite the extreme difference in correlation from the pair-wise comparisons involving enhanced CMAP and the comparisons involving standard CMAP, they exhibit similar trends. For example, in 1990 and 1998 all of the pair-wise comparisons had a lowered correlation, except for GPCP-APHRODITE (Fig. 33).
Moving higher in elevation, to the mountainous high lands, the correlations in the dry season are most varied. Again, the correlations including the standard and enhanced CMAP dataset are the lowest. All of the pair-wise comparisons have a lower correlation except for those involving the enhanced CMAP dataset. The correlation of pair-wise comparisons including the enhanced CMAP dataset are still ranging in values from -0.4 to 0.2. These correlation values are very similar to the enhanced CMAP correlation values from the other elevations in the dry season (Fig. 33 and 34). Also, it is interesting to note that the correlations that involve the enhanced CMAP dataset for the high lands during the dry season reflect many of the same trends in changes in correlation values over time. For example, from 1986 to 1995, the year to year increases and decreases in correlation between the pair-wise comparisons are seen in similar changes by the correlations including the enhanced CMAP dataset (Fig. 34).
The anomaly correlations during the monsoon season have very different characteristics than the correlations during the dry season. During the monsoon season in the lower elevations, the correlation values in 1984 for the pair-wise comparisons including standard CMAP (Fig. 35) are not as low as the same correlations during the dry season in 1984 (Fig. 32). Also, the correlations including the enhanced CMAP dataset are of interest. There is a much wider range of correlation values (-0.3 to 0.6) from correlations involving the enhanced CMAP dataset during the monsoon season, then in the dry season (Fig. 35 and 32). It is a common trait for monsoon anomaly correlation values for pair-wise comparisons including enhanced CMAP to have wide ranges. These wide ranges are seen again in the transitional (-0.5 to 0.6) and mountainous (-0.3 to 0.6) elevations (Fig. 36 and 37).
The trends of the correlations are also mimicked in many of the pair-wise comparisons. For example, in Fig. 36 the correlations including the enhanced CMAP dataset and the standard CMAP dataset have similar decreases and increases in correlation values with other datasets in 1987 to 1990, and again in 1997. As the elevation has increased, the variability of correlations involving enhanced CMAP have a wider range (-0.5 to 0.6). It is interesting to note, that from 1981 to 1989, the correlations involving both the standard and enhanced CMAP datasets are showing similar trends and similar values of correlation. After 1989, the value of the correlations for standard and enhanced CMAP datasets begin to diverge, with correlations involving standard CMAP increasing in value, and correlation including the enhanced CMAP remaining closer to a value of zero.

Figure 35: Low Lands, Monsoon Season Annual Anomaly Correlation
As the elevations increase in elevation, the usually significantly lower correlation values that include the enhanced CMAP become more similar to the range of the correlation values of the other datasets. This is because the correlation values of the other pair-wise comparisons are lowered with increased elevation while the correlation values involving enhanced CMAP are increased. The reanalysis data present in the enhanced CMAP data causes the increase in correlation during the monsoon season (Fig. 37). However, these correlation values are still the lowest of all of the pair-wise comparisons, except for in 1992 (enhanced CMAP – CHIRPS has a correlation of nearly 0.6, which is higher than the GPCP-standard CMAP and standard CMAP-CHIRPS correlations) and 2001 (enhanced CMAP-APHRODITE is higher than standard CMAP-APHRODITE).

Figure 36: Transitional Elevations, Monsoon Season Annual Anomaly Correlation
Figure 37: Mountainous, Monsoon Season Annual Anomaly Correlation
CHAPTER 6

CONCLUSION AND DISCUSSION

A. Key Findings

The goal of this study was to generate measures of correlation between climatic precipitation datasets, in particular GPCP, CMAP, CHIRPS, and APHRODITE in the GBM basin. This goal was obtained through meeting each of the four objectives.

The first object was “to find a broad temporal and spatial metric to determine which datasets are reporting greater or lesser precipitation over the entire study area and during the entire study period.” This was accomplished by the use of the EDF (Fig. 15 and 16). Here, it was concluded that CHIRPS and APHRODITE are reporting more rainfall events that are closer to average (represented by zero on the x-axis), than GPCP and CMAP. In other words, GPCP and CMAP are reporting more extreme positive and negative anomalies. This result is also indicative of the manner in which the data was handled. This may be due to the ability of the passive microwave scattering to capture intense rainfall events compared to the rain gauges with have
sparser, less uniform coverage. Since the higher resolution APHRODITE and CHIRPS data were averaged to 2.5° resolution, many of the individual cells that recorded localized rainfall events, thus CHIRPS and APHRODITE recorded higher and lower rainfall anomalies than GPCP and CMAP. However, the CHIRPS and APHRODITE maximum and minimum values were averaged out because there were more cells that reported no rain.

Objective 1 was also addressed by the spatial correlation maps (Fig. 17) and the graphs over time (Fig. 18). These maps showed that the geographic locations play a role in affecting the correlation of the pairs of datasets; in particular it had a major effect on any correlation that paired a dataset with CMAP in mountainous and transition elevations. While the spatial maps (Fig. 17) provided a method of looking at spatial correlation differences, the graph of the correlations over time (Fig. 18) was useful to show how correlations change over time with varying changes in technology used in the merging technique (i.e. number of stations, changes in satellite instruments, etc.). From this graph, there were significantly lower correlation values in 1984. This decrease came during a year when there was a drastic reduction in GPCC rain gauge data, causing a decrease in correlations including CMAP.

Finally, the first objective was met by graphing the monthly correlation of datasets (Fig. 19). This graph illustrated a clear decrease in correlations of all pair-wise comparisons during the monsoon months, emphasizing the importance of recognizing the different seasons in climate precipitation studies.

The second objective was “to find a correlation that will describe agreement and disagreement at different elevation zones.” The elevation zones were created based on the classification of the average elevation from the DEM from HydroSHEDS for each cell of the
grid over the study area. There were three classes created: Low lands (0 to 464 meters), Transitional elevations (465 to 1300 meters), and Mountainous elevations (1301 to 5000 meters). Furthermore, in the spatial maps (Fig. 17), it became apparent that topography plays a major role in affecting the correlation of pair-wise comparisons. In Table 3 and 4, the Spearman Rank Correlation was used to illustrate that the difference in correlation values at different elevations zones. Here it was shown that the pair-wise comparisons between APHRODITE and the other dataset are near zero in mountainous regions, implying that APHRODITE’s data is nearly independent from the other datasets in mountainous regions.

In figures 32 through 37, the Anomaly Correlation of every pair-wise comparison for each of the three elevation zones were compared. From these figures, it was found that enhanced CMAP was the poorest pair-wise comparison partner. All datasets compared to enhanced CMAP had lower correlations, then when compared to other datasets. This is likely due to the enhanced CMAP merging technique which includes NCEP/NCAR reanalysis data. This model data will tend to report precipitation even if there was no actual precipitation that day. Because of this overestimate, when NCEP/NCAR reanalysis data was used, it increased the correlation in the monsoon season. The correlation of GPCP and APHRODITE was the highest, suggesting GPCP’s use of multiple satellite inputs and APHRODITE’s dense network of rain gauges are reporting similar amounts of precipitation. This is a key finding since GPCP uses a merged satellite and rain gauge product, while APHRODITE uses only in-situ rain gauge data, thus implying that the values of precipitation from the merged product of GPCP are very similar to the actual amount of rainfall being recorded at the rain gauges. The correlations involving CHIRPS seem to result in values that are lower than the GPCP and APHRODITE pair-wise comparisons but higher than the standard and enhanced CMAP correlation values. The
correlation values for pair-wise comparisons including CHIRPS are lowered because of the lower density of CHIRPS rain gauges in this region, but the merging of infrared satellite data kept the correlation values higher than those that include either version of CMAP, which are very dependent on rain gauge data. Across all of these comparisons, the correlation values decreased as the elevation increased, except for the pair-wise comparisons including enhanced CMAP.

The third objective was “to find a correlation that will describe agreement and disagreement during the dry and monsoon season.” This objective was met by using several of the same figures that were used to address objective 2. In Tables 3 and 4, the Spearman Rank Correlation values for all of the pair-wise comparisons are higher in the Dry season then they are in the Monsoon season. This suggests that all datasets are in better agreement with determining when there is not rain, than determining when there is rain. The Anomaly Correlation as shown in Figures 32 to 37 show very similar results in respect to comparison of the dry and monsoon season; the pair-wise comparisons have a higher correlation during the dry season than the monsoon season, except for the pair-wise comparisons including the enhanced CMAP. Since the Anomaly Correlation is graphed over time, it is interesting to note that the correlation values increase for both the dry season and monsoon season over time for all of the comparisons except for those including enhanced CMAP. This implies that the input source data from each merged dataset are likely agreeing more on rainfall with time, except for the reanalysis data used in the enhanced CMAP data.

The fourth objective was “to determine if the density of rain gauges has affected the correlation between the datasets.” This was explored by Figures 20 through 31 and the graphs in Appendix B. Correlations involving CMAP, APHRODITE, and CHIRPS seem to be much more influenced by the presence and density of rain gauges than GPCP, although it was clear from the
graphs that there were other influences on the anomaly correlations than just the rain gauges. One possible reason for this is that there is no reason to expect the rainfall anomaly spatial structure to correlate with differences in gauge density patterns among the data sets. This may introduce significant noise among the different data sets with regard to their average gauge density.

B. Recommendations

First, there are recommendations relevant to end-users based on the findings. The findings presented in this study show that APHRODITE and GPCP have the highest correlation throughout the study period and area except for mountainous elevations during the monsoon season, as shown by the Spearman’s Rank Correlation. Furthermore, as seen in the Spearman’s Rank Correlations and Fig. 37, correlations involving CHIRPS are typically higher than GPCP-APHRODITE during periods when the correlations of GPCP-APHRODITE are low during the monsoon season in the mountainous elevations (see 1983, 1986, 1991, 2002, and 2006 on Fig. 37 for examples of when correlations including CHIRPS were higher than the correlation value of GPCP-APHRODITE). A merged product of APHRODITE, GPCP, and CHIRPS would be recommended to end-users and decision makers looking for a climatic precipitation dataset. The product could be constructed by weighting the precipitation values of individual 0.5° x 0.5° cells of APHRODITE at daily time steps. The weighting would be based on the monthly average anomaly value of the GPCP cell that the APHRODITE cell lies within. This will allow end-users looking for data for hydrology models to identify the beginning and end of the monsoon season, identify the intensity of the monsoon season, and determine frequency of precipitation in general. However, these weighting values would be used for elevations other than those that are classified
as mountainous. In these instances, monthly CHIRPS cells are averaged to 0.5° x 0.5° to match the resolution of APHRODITE. These CHIRPS averaged cells would be used to weight the daily cell values of APHRODITE precipitation in mountainous regions.

Next, there are several recommendations for end-users considering the use of any of the four merged datasets. The disagreement between APHRODITE and the other datasets during the monsoon season is a result of the merging technique for APHRODITE. APHRODITE weights cell values across mountainous regions to correct for the rain shadow effect and in doing so; the weighting technique may at times incorrectly adjust the rain fall amounts. Another recommendation is that the weighting function of APHRODITE be reconsidered for the extremely mountainous region, like those in the GBM basin.

APHRODITE’s through collection of rain gauge data has provided it with an excellent representation of precipitation over the region at a high resolution of 0.5°. This is a fine enough resolution so that mesoscale climatology could be effectively mapped and studied. Also, since APHRODITE is offered at daily time steps, the frequency, duration, and intensity of the precipitation can be studied for future hydrology research. APHRODITE would likely perform well with hydrology and agriculture models, though it should be studied to see if any significant differences exist when GPCP and CHRIPS are used as input data for these models as well.

GPCP may not have the high spatial resolution of APHRDOITE, but given its high agreement with APHRODITE, it is a good representation of the synoptic scale climatology of the GBM Basin. Given GPCP’s high correlation values with APHRODITE, it can be concluded that the merging technique of GPCP creates a product that is very similar to the precipitation being reported by the rain gauges. However, given its coarse spatial (2.5°) and temporal (monthly)
resolution, GPCP would require more processing to be able to provide a high enough spatial or temporal resolution to accurately describe the intensity, duration, and frequency of precipitation.

CHIRPS offers the highest resolution of the four datasets, but it does strongly agree or disagree with the other datasets (Figures 20 through 25). In the GBM basin, CHIRPS has a relatively low amount of rain gauges available, causing its correlation to other datasets to be lowered. However, CHIRPS’ use of infrared satellite data and high climatology datasets keeps CHIRPS in a moderate agreement with the other datasets. CHIRPS might perform better when compared to higher resolution datasets, because of possible loss of data as 2,500 cells of CHIRPS data has to be averaged for each cell of the grid used in this study. CHIRPS needs further work comparing it to other higher resolution datasets, but could be used to identify and describe localized climates within the GBM Basin in the Low lands during the dry and monsoon season.

CMAP is the most problematic dataset because of its disagreement with the other datasets. While CMAP uses many of the same source data as GPCP, its merging technique is different enough from GPCP that it results in very different recorded anomalies in this region. While other studies have found that GPCP and CMAP correlate very well with one another on a global scale (Yin et al, 2004), CMAP does not correlate well with GPCP in the GBM Basin. CMAP, and the other datasets, could benefit from an increase in rain gauge data or a change in source IR data. Further, the use of the enhanced CMAP dataset is very discouraged given its nearly independent relation to the other datasets.

Although it is outside of the scope of this research, another recommendation for decision makers is to consider installing more rain gauge stations. Their data could be incorporated into each of the datasets to better improve correlations in mountainous areas. It should be restated that
the data in this study focused on the boundaries of the GBM Basin and not political boundaries. This could provide multiple challenges for countries to work together to making a denser, more robust rain gauge network. However, it is a challenge worth taking because countries share the same basin and thus much of the same water. With more rain gauge stations installed, data could be provided for water resource management, providing more information about the complicated relationship between the extreme mountainous regions of the Himalayas and the South Asia monsoon. In order to help bridge the political boundaries, making contact and sharing these recommendations with the International Center for Integrated Mountain Development (ICIMOD), or its global partner the Regional Visualization and Monitoring System (SERVIR).

C. Future Work

There are numerous studies that could be conducted to further this comparative study. Using APHRODITE, CHIRPS, and GPCP as a merged input for a hydrology model, like the Variable Infiltration Capacity (VIC) hydrology model, the difference from the modeled flow and the actual flow could be calculated to determine if the new merged data addresses the disagreements listed above. These results would be shared with ICIMOD and other local decision makers for their water management services.

Another type of model that could use an APHRODITE, CHIRPS, and GPCP merged product are agricultural crop yield models, such as the Decision Support System for Agrotechnology Transfer (DSSAT). Models such as these need a precipitation and climate to
simulate crops yield. Again, the new merged dataset could be used to generate a modeled crop yield that could be compared to the actual yield from previous years.

An even further step would be to take the information from this current study, the hydrology study, and the agriculture study and construct a vulnerability index for the people of the GBM basin. This index would also have to take into account other factors such as temperature, natural disasters, and social and demographic information. This could provide essential information for the people and decision makers of the GBM Basin as the threat of climate change approaches.
REFERENCES

ICIMOD. N. p., 24 June 2013.


APPENDIX A

Python Scripts

The general methodology for each dataset, except for CHIRPS since the data was available as a GeoTIFF, was to first write a script that would open the NetCDF and write the data to an array. The second script was written to clip, flip the orientation (if needed), and run a zonal statistics tool that would generate a table of averages for each cell. While there were minor differences between each sets of code, only one example of each will be shown for brevity.

GPCP_corrected.py

#This script will open the NetCDF, and write out the array as an ASCII Grid File so that it can be #treated as a raster with the arcpy toolkit.

import scipy
from scipy.io import netcdf
import numpy
import glob
import arcpy

f = netcdf.netcdf_file("C:\Users\James\Documents\Thesis\GCP\precip.mon.mean.nc")
fGPCP = f.variables['precip']
f.close()}
#JAN
Jan_1981=[]
Jan_1982=[]
Jan_1983=[]
Jan_1984=[]
Jan_1985=[]
Jan_1986=[]
Jan_1987=[]
Jan_1988=[]
Jan_1989=[]
Jan_1990=[]
Jan_1991=[]
Jan_1992=[]
Jan_1993=[]
Jan_1994=[]
Jan_1995=[]
Jan_1996=[]
Jan_1997=[]
Jan_1998=[]
Jan_1999=[]
Jan_2000=[]
Jan_2001=[]
Jan_2002=[]
Jan_2003=[]
Jan_2004=[]
Jan_2005=[]
Jan_2006=[]
Jan_2007=[]
Jan_2008=[]
Jan_2009=[]
Jan_2010=[]

Jan_1981=fGPCP[24,:,::]
Jan_1982=fGPCP[36,:,::]
Jan_1983=fGPCP[48,:,::]
Jan_1984=fGPCP[60,:,::]
Jan_1985=fGPCP[72,:,::]
Jan_1986=fGPCP[84,:,::]
Jan_1987=fGPCP[96,:,::]
Jan_1988=fGPCP[108,:,::]
Jan_1989=fGPCP[120,:,::]
Jan_1990=fGPCP[132,:,::]
Jan_1991=fGPCP[144,:,::]
Jan_1992=fGPCP[156,:,::]
Jan_1993=fGPCP[168,:,::]
Jan_1994=fGPCP[180,:,::]
Jan_1995=fGPCP[192,:,::]
Jan_1996=fGPCP[204,:,::]
Jan_1997=fGPCP[216,:,::]
Jan_1998=fGPCP[228,:,::]
Jan_1999=fGPCP[240,:,::]
Jan_2000=fGPCP[252,:,::]
Jan_2001=fGPCP[264,:,::]
Jan_2002=fGPCP[276,:,::]
Jan_2003=fGPCP[288,:,:]
Jan_2004=fGPCP[300,:,:]
Jan_2005=fGPCP[312,:,:]
Jan_2006=fGPCP[324,:,:]
Jan_2007=fGPCP[336,:,:]
Jan_2008=fGPCP[348,:,:]
Jan_2009=fGPCP[360,:,:]
Jan_2010=fGPCP[372,:,:]


Jan_path = ['C:\Users\James\Documents\Thesis\GPCP\Jan_1981.asc','C:\Users\James\Documents\Thesis\GPCP\Jan_1982.asc','C:\Users\James\Documents\Thesis\GPCP\Jan_1983.asc','C:\Users\James\Documents\Thesis\GPCP\Jan_1984.asc','C:\Users\James\Documents\Thesis\GPCP\Jan_1985.asc','C:\Users\James\Documents\Thesis\GPCP\Jan_1986.asc','C:\Users\James\Documents\Thesis\GPCP\Jan_1987.asc','C:\Users\James\Documents\Thesis\GPCP\Jan_1988.asc','C:\Users\James\Documents\Thesis\GPCP\Jan_1989.asc','C:\Users\James\Documents\Thesis\GPCP\Jan_1990.asc','C:\Users\James\Documents\Thesis\GPCP\Jan_1991.asc','C:\Users\James\Documents\Thesis\GPCP\Jan_1992.asc','C:\Users\James\Documents\Thesis\GPCP\Jan_1993.asc','C:\Users\James\Documents\Thesis\GPCP\Jan_1994.asc','C:\Users\James\Documents\Thesis\GPCP\Jan_1995.asc','C:\Users\James\Documents\Thesis\GPCP\Jan_1996.asc','C:\Users\James\Documents\Thesis\GPCP\Jan_1997.asc','C:\Users\James\Documents\Thesis\GPCP\Jan_1998.asc',]
i=0
while (i <30):
    fname = open(Jan_path[i], 'w')
    fname.write("ncols 144\n")
    fname.write("nrows 72\n")
    fname.write("xllcorner 0.000\n")
    fname.write("yllcorner -90.000\n")
    fname.write("cellsize 2.5\n")
    fname.write("NODATA_value -9999\n")
    for precip_rec in Jan_years[i]:
        fname.write(str(precip_rec) + '\n')
    fname.close()
    i=i+1

#This procedure was repeated for each month. Similar codes were written for CMAP and
#APHRODITE, though APHRODITE had to calculate monthly averages of daily rainfall before
#the climatologies or anomalies could be calculated.
This script calculated the monthly climatology and the monthly anomalies and then finally created a table of statistics for each cell of the study grid. Similar scripts were written for CMAP, APHRODITE, and CHIRPS.

```python
import glob
import arcpy
from arcpy.sa import *

arcpy.CheckOutExtension("Spatial")

ASCII_Ras = glob.glob("C:\Users\James\Documents\Thesis\GPCP\*.asc")

for ASCII in ASCII_Ras:
    outras = ASCII[:-4] + ".tif"
    arcpy.Clip_management(ASCII, "72.5 20 100 32.5", outras,"E:\James_Folders\Thesis\BasinShapeFiles\GBM_Basin_Grid.shp", "-9999", "ClippingGeometry")
    print "clipped " + outras

jan=[]
feb=[]
mar=[]
apr=[]
may=[]
jun=[]
 jul=[]
 aug=[]
sep=[]
 ocb=[]
nov=[]
dec=[]
```
ASCII_clip = glob.glob("C:\Users\James\Documents\Thesis\GCP\*_clip.tif")

for clip in ASCII_clip:
    print clip[37:40]
    if clip[37:40] == "Jan":
        jan.append(clip)
    elif clip[37:40] == "Feb":
        feb.append(clip)
    elif clip[37:40] == "Mar":
        mar.append(clip)
    elif clip[37:40] == "Apr":
        apr.append(clip)
    elif clip[37:40] == "May":
        may.append(clip)
    elif clip[37:40] == "Jun":
        jun.append(clip)
    elif clip[37:40] == "Jul":
        jul.append(clip)
    elif clip[37:40] == "Aug":
        aug.append(clip)
    elif clip[37:40] == "Sep":
        sep.append(clip)
    elif clip[37:40] == "Ocb":
        ocb.append(clip)
    elif clip[37:40] == "Nov":
        nov.append(clip)
    elif clip[37:40] == "Dec":
        dec.append(clip)
#JAN

jan_81 = jan[0]
jan_82 = jan[1]
jan_83 = jan[2]
jan_84 = jan[3]
jan_85 = jan[4]
jan_86 = jan[5]
jan_87 = jan[6]
jan_88 = jan[7]
jan_89 = jan[8]
jan_90 = jan[9]
jan_91 = jan[10]
jan_92 = jan[11]
jan_93 = jan[12]
jan_94 = jan[13]
jan_95 = jan[14]
jan_96 = jan[15]
jan_97 = jan[16]
jan_98 = jan[17]
jan_99 = jan[18]
jan_00 = jan[19]
jan_01 = jan[20]
jan_02 = jan[21]
jan_03 = jan[22]
jan_04 = jan[23]
jan_05 = jan[24]
\[
\begin{align*}
\text{jan}_06 &= \text{jan}[25] \\
\text{jan}_07 &= \text{jan}[26] \\
\text{jan}_08 &= \text{jan}[27] \\
\text{jan}_09 &= \text{jan}[28] \\
\text{jan}_10 &= \text{jan}[29] \\
\text{jan}_{81-82} &= \text{Plus}(\text{jan}_81, \text{jan}_82) \\
\text{jan}_{83-84} &= \text{Plus}(\text{jan}_83, \text{jan}_84) \\
\text{jan}_{85-86} &= \text{Plus}(\text{jan}_85, \text{jan}_86) \\
\text{jan}_{87-88} &= \text{Plus}(\text{jan}_87, \text{jan}_88) \\
\text{jan}_{89-90} &= \text{Plus}(\text{jan}_89, \text{jan}_90) \\
\text{jan}_{91-92} &= \text{Plus}(\text{jan}_91, \text{jan}_92) \\
\text{jan}_{93-94} &= \text{Plus}(\text{jan}_93, \text{jan}_94) \\
\text{jan}_{95-96} &= \text{Plus}(\text{jan}_95, \text{jan}_96) \\
\text{jan}_{97-98} &= \text{Plus}(\text{jan}_97, \text{jan}_98) \\
\text{jan}_{99-00} &= \text{Plus}(\text{jan}_99, \text{jan}_00) \\
\text{jan}_{01-02} &= \text{Plus}(\text{jan}_01, \text{jan}_02) \\
\text{jan}_{03-04} &= \text{Plus}(\text{jan}_03, \text{jan}_04) \\
\text{jan}_{05-06} &= \text{Plus}(\text{jan}_05, \text{jan}_06) \\
\text{jan}_{07-08} &= \text{Plus}(\text{jan}_07, \text{jan}_08) \\
\text{jan}_{09-10} &= \text{Plus}(\text{jan}_09, \text{jan}_10) \\
\text{jan}_{81-84} &= \text{Plus}(\text{jan}_{81-82}, \text{jan}_{83-84}) \\
\text{jan}_{85-88} &= \text{Plus}(\text{jan}_{85-86}, \text{jan}_{87-88}) \\
\text{jan}_{89-92} &= \text{Plus}(\text{jan}_{89-90}, \text{jan}_{91-92}) \\
\text{jan}_{93-96} &= \text{Plus}(\text{jan}_{93-94}, \text{jan}_{95-96}) \\
\text{jan}_{97-00} &= \text{Plus}(\text{jan}_{97-98}, \text{jan}_{99-00})
\end{align*}
\]
\[
\begin{align*}
\text{jan} \_01 \_04 &= \text{Plus}(\text{jan} \_01 \_02, \text{jan} \_03 \_04) \\
\text{jan} \_05 \_08 &= \text{Plus}(\text{jan} \_05 \_06, \text{jan} \_07 \_08) \\
\text{jan} \_05 \_10 &= \text{Plus}(\text{jan} \_05 \_08, \text{jan} \_09 \_10) \\
\text{jan} \_81 \_88 &= \text{Plus}(\text{jan} \_81 \_84, \text{jan} \_85 \_88) \\
\text{jan} \_89 \_96 &= \text{Plus}(\text{jan} \_89 \_92, \text{jan} \_93 \_96) \\
\text{jan} \_97 \_04 &= \text{Plus}(\text{jan} \_97 \_00, \text{jan} \_01 \_04) \\
\text{jan} \_81 \_96 &= \text{Plus}(\text{jan} \_81 \_88, \text{jan} \_89 \_96) \\
\text{jan} \_97 \_10 &= \text{Plus}(\text{jan} \_97 \_04, \text{jan} \_05 \_10) \\
\text{jan} \_81 \_10 &= \text{Plus}(\text{jan} \_81 \_96, \text{jan} \_97 \_10) \\
\text{jan} \_81 \_10 &= \text{Divide}(\text{jan} \_81 \_10, 30.0000) \\
\text{jan} \_81 \_season \_rmv &= \text{Minus}(\text{jan} \_81, \text{jan} \_avg) \\
\text{jan} \_81 \_season \_rmv \text{.save}("C:\Users\James\Documents\Thesis\GPCP\jan\_81\_season\_rmv.tif") \\
\text{jan} \_82 \_season \_rmv &= \text{Minus}(\text{jan} \_82, \text{jan} \_avg) \\
\text{jan} \_82 \_season \_rmv \text{.save}("C:\Users\James\Documents\Thesis\GPCP\jan\_82\_season\_rmv.tif") \\
\text{jan} \_83 \_season \_rmv &= \text{Minus}(\text{jan} \_83, \text{jan} \_avg) \\
\text{jan} \_83 \_season \_rmv \text{.save}("C:\Users\James\Documents\Thesis\GPCP\jan\_83\_season\_rmv.tif") \\
\text{jan} \_84 \_season \_rmv &= \text{Minus}(\text{jan} \_84, \text{jan} \_avg) \\
\text{jan} \_84 \_season \_rmv \text{.save}("C:\Users\James\Documents\Thesis\GPCP\jan\_84\_season\_rmv.tif") \\
\text{jan} \_85 \_season \_rmv &= \text{Minus}(\text{jan} \_85, \text{jan} \_avg)
\end{align*}
\]
The code snippet provided calculates the seasonal anomalies for each January from 1985 to 1995 and saves them to a file. Here is the code in a more readable format:

```python
jan_85_season_rmv = Minus(jan_85, jan_avg)
jan_85_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_85_season_rmv.tif")

jan_86_season_rmv = Minus(jan_86, jan_avg)
jan_86_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_86_season_rmv.tif")

jan_87_season_rmv = Minus(jan_87, jan_avg)
jan_87_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_87_season_rmv.tif")

jan_88_season_rmv = Minus(jan_88, jan_avg)
jan_88_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_88_season_rmv.tif")

jan_89_season_rmv = Minus(jan_89, jan_avg)
jan_89_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_89_season_rmv.tif")

jan_90_season_rmv = Minus(jan_90, jan_avg)
jan_90_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_90_season_rmv.tif")

jan_91_season_rmv = Minus(jan_91, jan_avg)
jan_91_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_91_season_rmv.tif")

jan_92_season_rmv = Minus(jan_92, jan_avg)
jan_92_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_92_season_rmv.tif")

jan_93_season_rmv = Minus(jan_93, jan_avg)
jan_93_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_93_season_rmv.tif")

jan_94_season_rmv = Minus(jan_94, jan_avg)
jan_94_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_94_season_rmv.tif")

jan_95_season_rmv = Minus(jan_95, jan_avg)
jan_95_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_95_season_rmv.tif")
``
jan_96_season_rmv = Minus(jan_96, jan_avg)
jan_96_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_96_season_rmv.tif")

jan_97_season_rmv = Minus(jan_97, jan_avg)
jan_97_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_97_season_rmv.tif")

jan_98_season_rmv = Minus(jan_98, jan_avg)
jan_98_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_98_season_rmv.tif")

jan_99_season_rmv = Minus(jan_99, jan_avg)
jan_99_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_99_season_rmv.tif")

jan_00_season_rmv = Minus(jan_00, jan_avg)
jan_00_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_00_season_rmv.tif")

jan_01_season_rmv = Minus(jan_01, jan_avg)
jan_01_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_01_season_rmv.tif")

jan_02_season_rmv = Minus(jan_02, jan_avg)
jan_02_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_02_season_rmv.tif")

jan_03_season_rmv = Minus(jan_03, jan_avg)
jan_03_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_03_season_rmv.tif")

jan_04_season_rmv = Minus(jan_04, jan_avg)
jan_04_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_04_season_rmv.tif")

jan_05_season_rmv = Minus(jan_05, jan_avg)
jan_05_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_05_season_rmv.tif")

jan_06_season_rmv = Minus(jan_06, jan_avg)
jan_06_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_06_season_rmv.tif")

jan_07_season_rmv = Minus(jan_07, jan_avg)
jan_07_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_07_season_rmv.tif")

jan_08_season_rmv = Minus(jan_08, jan_avg)
jan_08_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_08_season_rmv.tif")

jan_09_season_rmv = Minus(jan_09, jan_avg)
jan_09_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_09_season_rmv.tif")

jan_10_season_rmv = Minus(jan_10, jan_avg)
jan_10_season_rmv.save("C:\Users\James\Documents\Thesis\GPCP\jan_10_season_rmv.tif")

#This was repeated for each month.

Tif_season_rmv = glob.glob("C:\Users\James\Documents\Thesis\GPCP\*season_rmv.tif")

for tif in Tif_season_rmv:
    outTable= tif[4:]+".dbf"

    ZonalStatisticsAsTable("E:\James_Folders\Thesis\BasinShapeFiles\GBM_Basin_Grid.shp", "FID", tif, outTable, "NODATA", "MEAN")
    print "Created table for " +tif

Similar codes were written to get the station density for each dataset so that the number of stations was saved as a raster. Below is the code used for APHRODITE to map the rain gauge density for one month of one year.

Station_short.py

import scipy
from scipy.io import netcdf
year_STN_1981 = netcdf.netcdf_file('C:\Users\James\Documents\Thesis\APHRODITE\25Degree\APHRO_MA_025deg_V1101.1981.nc', 'r')

stn = year_STN_1981.variables['rstn']
year_STN_1981.close()

STN_1981_jan=0
STN_1981_jan=stn[0,:,:]+STN_1981_jan
STN_1981_jan=stn[1,:,:]+STN_1981_jan
STN_1981_jan=stn[4,:,:]+STN_1981_jan
STN_1981_jan=stn[5,:,:]+STN_1981_jan
STN_1981_jan=stn[6,:,:]+STN_1981_jan
STN_1981_jan=stn[7,:,:]+STN_1981_jan
STN_1981_jan=stn[8,:,:]+STN_1981_jan
STN_1981_jan=stn[9,:,:]+STN_1981_jan
STN_1981_jan=stn[10,:,:]+STN_1981_jan
STN_1981_jan=stn[12,:,:]+STN_1981_jan
STN_1981_jan=stn[14,:,:]+STN_1981_jan
STN_1981_jan=stn[15,:,:]+STN_1981_jan
STN_1981_jan=stn[16,:,:]+STN_1981_jan
STN_1981_jan=stn[17,:,:]+STN_1981_jan
STN_1981_jan=stn[18,:,:]+STN_1981_jan
STN_1981_jan=stn[19,:,:]+STN_1981_jan
STN_1981_jan=stn[20,:,:]+STN_1981_jan
STN_1981_jan=stn[21,:,:]+STN_1981_jan
STN_1981_jan=stn[22,:,:]+STN_1981_jan
STN_1981_jan=stn[23,:,:]+STN_1981_jan
STN_1981_jan=stn[26,:,:]+STN_1981_jan
STN_1981_jan=stn[27,:,:]+STN_1981_jan
STN_1981_jan=stn[29,:,:]+STN_1981_jan
STN_1981_jan=stn[30,:,:]+STN_1981_jan
STN_1981_jan=STN_1981_jan/31.000

fname=
open('C:\Users\James\Documents\Thesis\APHRODITE\25Degree\station\monthly\stn_1981_jan.asc','w')
fname.write("ncols 360 \n")
fname.write("nrows 280 \n")
fname.write("xllcorner 60 \n")
fname.write("yllcorner -15 \n")
fname.write("cellsize 0.25 \n")
fname.write("NODATA_value -99.9 \n")
for stn_density in STN_1981_jan: fname.write(str(stn_density)+\n")
fname.close()
APPENDIX B

Graphs of station density and Anomaly Correlations

![Graphs of station density and Anomaly Correlations](image)