Understanding Climate Change and Variability III: A Spatio-Temporal Data Mining Perspective

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

James Hocine Faghmous

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor Of Philosophy

Vipin Kumar

May, 2013

In your Name, I write.

”And they do not encompass anything of His Knowledge, except whatever He wills”

Acknowledgements

Although this thesis lists a single author, it is the culmination of a collaborative effort involving more people than this acknowledgement could fit. I would like to gratefully thank my advisor, Professor Vipin Kumar, for consistently pushing me to be a better researcher and for his unconditional support, guidance, and optimism. Not a single day working with you felt like "work" (except, when filing for travel reimbursement). I also thank my thesis committee Professors Daniel Boley, Arindam Banerjee, and Snigdhansu Chaterjee for their service and counsel. I would like to thank my family who sacrificed everything to give me every opportunity to be a better person. Specifically, I thank my dear mother Margaret Dwyer, for being the model of dedication, focus, kindness, and integrity that I strive to emulate daily. I thank my step-father, Ethan Pirk, for loving me like a son, helping me be a better son to my mother, and for braving air-travel, twice, to be by my side during important chapters in my life. I thank my aunt, Yamina Faghmous, for providing me with all the love and support a young boy could ask for and for teaching me how to sew at age 6. I thank my father, Mohammed Faghmous, for his love and sacrifice, and for teaching me to "always do the right thing" regardless of how inconvenient that might be. Finally, I thank my entire family in Algeria and my wonderful parents and siblings in-law who love me despite all my shortcomings.

Friendship has been an essential part of my support system and without the love and support of my friends I would not have been able to get through this demanding chapter of my life. I would like to thank my childhood best friend, Amir Megdoud, for giving me the courage to have bigger ambitions. I thank my two college best friends, Nezar Mouaki-Banani and Hakim Bouchendoukha, for their unconditional friendship and endless studying sessions. I would like to thank my great friend Franklin Khaleel Cromer and his family for their love, support, and guidance. I also thank my wonderful

Minnesota friends who made my six and half years here filled with wonderful memories. I especially thank the Zain, Yousif, El-Sawaf and Hannon families. I would like to also thank my co-founders at MuslimBuddy, Inc., Muneer Karcher-Ramos, Sami Khwaja, Naaima Khan, Muaz Rushdi, and Khaled El-Sawaf for expanding my thinking, providing me with an outlet to balance my academic life, and allowing me to serve my community. I would also like to thank my former teammates and coaches in Algeria and at The City College of New York (CCNY) for countless hours of competition, camaraderie, and self-improvement.

I believe that ideas do not originate in a bubble and the ideas brought fourth in this thesis are, indeed, the result of years of collaborative learning, discussions, and analysis. I would like to thank Professor Susan Besse for radically extending my intellectual comfort zone and for entrusting me with my first research opportunity. I thank Robin Villa, Lee Linde, the CCNY Honors College staff for their support and encouragement through the grueling process of applying for graduate school and graduate funding. I thank Professor George Wolberg for instilling in me a "can do" mentality and the entrepreneurial bug. I would like to thank my late friend and mentor Professor Daniel D. McCracken who taught me that it is never too late to teach yourself something new. It is my hope that I will serve his memory well as I follow his footsteps as a computer science educator. I am thankful to my two doctoral classmates Dr. Muhammad A. Ahmad and Dr. Mohammed El-Idrisi for providing me with different perspectives on computer science research. I thank all my co-authors and collaborators for their support and stimulating discussions, especially: Professor Paul Schrater, Dr. Michel dos Santos Mesquita, Dr. Frode Vikebø, Dr. Shyam Boriah, Dr. Stefan Liess, Yashu Chamber, Varun Mithal, and Professor Kerry Emanuel. I also thank all my undergraduate research interns from whom I have learned tremendously: Ryan Haasken, Graham Smith, Luke Styles, Nikai Gibson, Matthew Le, Mohammed Uluyol, and Matthew Papke. I would like to thank the many administrators without whom this very document would not be in front of you. Specifically: Kathleen Clinton, Georganne Tolaas, Jennifer Olk, Liz Freppert, Peggy Stewart, and Laura Connor. Finally, I thank all the peer-reviewers and anyone with whom I exchanged ideas as they have helped improved many of the concepts in this thesis.

I would like to thank the National Institutes of Health (NIH) and the National

Science Foundation (NSF) for generously supporting my graduate research through a Neuro-Physical-Computational Graduate Fellowship and a Graduate Research Fellowship, respectively. I am also grateful for joint support from the NSF and the Norwegian National Research Council in the form of a Nordic Research Opportunity Fellowship, and finally I thank the University of Minnesota for a generous Doctoral Dissertation Fellowship. It is my hope that the research resulting from this generous funding will benefit one of the major contributors to my work: the American tax payer. I have been dedicated to using computer science to serve society, and I hope that I will honor your investment in my career.

Finally, I would like to thank my wife and serial co-founder, Zahra Aljabri, for her unwavering love, being a true partner in our marriage, and for the endless amounts of fun and joy she brings to my life. I am looking forward to dancing our lives away to the beat of our love.

Dedication

To my uncle, Abdul Majeed, for showing me that music and art have the power to lift the spirit well after the body has given up.

To my son, Mus'ab Adrien, may you pursue the dreams that are true in your heart yet inconceivable to your mind.

ABSTRACT

This thesis provides a computer science audience with an introduction to mining climate data with an emphasis on the singular characteristics of the datasets and research questions climate science attempts to address. We demonstrate some of the concepts discussed in the earlier parts of the thesis with two climate-related applications of relationship and pattern mining. In both instances, we show that insightfully mining the spatio-temporal context of climate datasets can yield significant improvements in the performance of learning algorithms. We focus on two spatio-temporal data mining applications one predicting Atlantic tropical cyclone (TC) activity and the other on mesoscale ocean eddy monitoring.

Tropical cyclones are among of the most devastating geophysical phenomena and predicting their occurrence has become a subject of intense scientific and societal interest. A large body of research focuses on using the large-scale environmental conditions to forecast cumulative TC activity on interannual scales. One of the known influencers of the large-scale conditions over the Atlantic ocean on interannual time-scales is the quasi-periodic warming and cooling cycle of the Pacific ocean, know as the El Niño Southern Oscillation (ENSO). Several research efforts have focused on capturing the ENSO cycle using empirical indices that average Pacific sea surface temperatures over fixed oceanic regions. These traditional indices have provided limited Atlantic TC forecasting insight, mainly because they have little predictive skill before the Northern Hemisphere spring (commonly known as the “spring predictability barrier”). We introduce the spatial ENSO index (S-ENSO) and show that it is better than traditional ENSO indices at forecasting Atlantic TC activity. Furthermore, S-ENSO is not susceptible to the ENSO spring predictability barrier traditional indices suffer from. Given the numerous global phenomena associated with ENSO, S-ENSO may be useful to researchers forecasting other phenomena associated with ENSO.

Our pattern mining application focuses on monitoring mesoscale ocean eddy dynamics. Mesoscale ocean eddies are coherent rotating structures that transport heat, salt, energy, and nutrients across oceans – also known as the “storms of the sea”. As a result, accurately identifying and tracking such phenomena are crucial for understanding ocean dynamics and marine ecosystem sustainability. This thesis proposes several advances to traditional eddy identification and monitoring algorithms. First, we are able to leverage the spatio-temporal context of the data to identify more physically-consistent features than existing methods. Second, we introduce a novel eddy tracking algorithm that not only resolves eddy tracks better than traditional methods, especially in the presence of noise but is also able to take corrective measures autonomously on the independent detection step. Finally, we provide the community with different evaluation metrics to assess the performance of unsupervised learning algorithms in the physical sciences.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Our world is experiencing simultaneous changes in population, industrialization, and climate amongst other planetary-scale changes. These contemporaneous transformations, known as *global change*, raise pressing questions of significant scientific and societal interest [103]. For example, how will the continued growth in global population and persisting tropical deforestation, or global climate change, affect our ability to access food and water? Coincidentally, these questions are emerging at a time when data, specifically spatio-temporal climate data, are more available than ever before. In fact, climate science promises to be one of the largest sources of data for data-driven research. A recent lower bound estimate puts the size of climate data in 2010 at 10 Petabyytes (1 PB = 1,000 TB). This number is projected to grow exponentially to about 350 Patabytes by 2030 [189].

The last decades have seen tremendous growth in data-driven learning algorithms and their broad-range applications [124]. This rapid growth was fueled by the Internet's democratization of data production, access, and sharing. Merely observing these events unfold – the growth of climate data, a wide-range of challenging real-world research questions, and the emergence of data mining and machine learning in virtually every domain where data are reasonably available – one may assume that data mining is ripe to make significant contributions to these challenges.

Unfortunately, this has not been the case – at least not at the scale we have come to expect from the success of data mining in other domains, such as biology and e-commerce. At a high level, this lack of progress is due to the inherent *nature* of climate data as well as the *types* of research questions climate science attempts to address.

Although the size of climate data is a serious challenge, there are major research efforts to address the variety, velocity, and volume of climate data (commonly referred to as Big Data's 3Vs). Research efforts to address the *nature* of climate data, however, are severely lagging the rate of data growth. For instance, climate data tend to be predominantly spatio-temporal, noisy, and heterogeneous. The spatio-temporal nature of climate data emerges in the form of auto- and cross-correlation between input variables. Therefore, existing learning methods that make implicit or explicit independence assumptions about the input data will have limited applicability to the climate domain.

It is also important to study the *types* of research questions that climate science brings forth. Climate science is the study of the spatial and temporal variations of the atmosphere-hydrosphere-land surface system over prolonged time periods. As a result, climate-related questions are inexorably linked to space and time. This means that climate scientists are interested in solutions that explain the evolution of phenomena in space and time. Furthermore, the majority of climate phenomena occur only within a specific region and time period. For example, hurricanes only take place in certain geographic regions and during a limited month range. However, due to the large datasets and the exponential number of space-time subsets within the data, we must reduce the complexity of problems by finding significant space-time subsets.

The combination of climate data's unique characteristics and associated research questions require the emergence of a new generation of space-time algorithms. Fortunately, climate data have intrinsic space and time information that, if insightfully leveraged, can provide a powerful computational framework to address many of the challenges listed above while significantly reducing the complexity of computational problems. In this thesis, we focus on the advances and opportunities for *spatio-temporal data mining*: a collection of methods that mine the data's spatio-temporal context to increase an algorithm's accuracy, scalability, or interpretability (relative to non-space-time aware algorithms). In the next section, we review this thesis' major contributions. The remaining of the thesis is organized as follows: Chapter 2 reviews the most common types of climate data and the associated challenges stemming from such datasets. Chapter 3 provides a broad overview of the types of STDM applications to climate data and a sample of works within each application. The rest of the thesis focuses on two

STDM applications: relationship and pattern mining. Chapter 4 reviews the state-of-the-art tropical cyclone forecasting methods and presents a spatio-temporal method to identify relationships between sea surface temperature and Atlantic tropical cyclones. Chapter 5 presents a spatio-temporal extension to commonly used static climate indices. Chapter 6 presents a spatial pattern mining method to autonomously identify features in continuous data. Chapter 7 presents a multiple object tracking algorithm that tracks the features identified in the previous chapter. Chapter 8 presents a temporal method to to autonomously identify features in continuous data. Finally, Chapter 9 reviews the major contributions of the thesis and presents research opportunities within STDM and climate.

1.1 Summary of contributions

This thesis makes several contributions to data-driven methods applied to climate challenges, specifically within the domains of spatio-temporal relationship and pattern mining. All the source code of the methods presented in this thesis is open-source and available for download at: `https://github.com/jfaghm`

Literature survey

We provide two major surveys in this thesis. The first provides a broad review of STDM application to climate data and related challenges. This survey provides the reader with the tools to make informed decisions about the techniques and opportunities available to analyze climate data. The second survey reviews the major works in tropical cyclone predictions with an emphasis on evaluation metrics and model performance. This survey provides the computer science researcher with a concise summary of the state-of-the-art in tropical cyclone predictions and discussions on opportunities to impact this active research field.

Scalable exploratory relationship mining

We introduce *QuickSig*, a suite of scalable methods to mine linear and nonlinear relationships between spatio-temporal variables at multiple spatio-temporal scales. Additionally, given the auto-correlated nature of the data, we provide Monte Carlo-based significance tests to asses the robustness of relationships discovered. These methods allow researchers to rapidly formulate and test hypotheses in large climate datasets.

Spatio-temporal analysis of climate data

We introduce *Spatial ENSO (S-ENSO)* the first spatio-temporal extension of the famous El-Niño Southern Oscillation index. Our work presents a simple and physically-based extension to static warming indices to account for the spatio-temporal warming patterns in the Pacific ocean. The spatio-temporal extension allows to predict tropical cyclone counts several months before traditional (non space-time aware) indices.

Spatio-temporal pattern mining of user-defined features in continuos data

We introduce two novel algorithms that are able to identify patterns in four-dimensional satellite data. This thesis presents and analyses two different, yet complimentary approaches to mining such patterns: one spatial and the other temporal. Both methods are able to significantly improve upon the state-of-the-art methods in terms of both accuracy and scalability. These results show that analyzing the spatio-temporal context of the data can yield significant improvements over traditional learning algorithms.

Unsupervised tracking of user-defined features and unsupervised self-learning using the data's spatio-temporal context

We present *Multiple Hypothesis Assignment (MHA)* a deferred-logic, multiple feature tracking algorithm. While MHA is able to track noisy features more accurately than traditional methods, it is also able to leverage the data's spatio-temporal context to autonomously identify errors within the features and correct them *a posteriori*. Additionally, MHA allows to study patterns such as merges and splits that are uncommon in the traditional object tracking literature.

Testing unsupervised learning methods

Finally, this thesis provides several evaluation procedures for learning algorithms dealing with unlabeled data. We demonstrate such procedures in the context of feature tracking in noisy and continuous data. These evaluation methods provide researchers with concrete examples on how to test unsupervised learning algorithms in the physical sciences.

Chapter 2

An Overview of Climate Data and Associated Challenges

2.0.1 Types of Climate Data

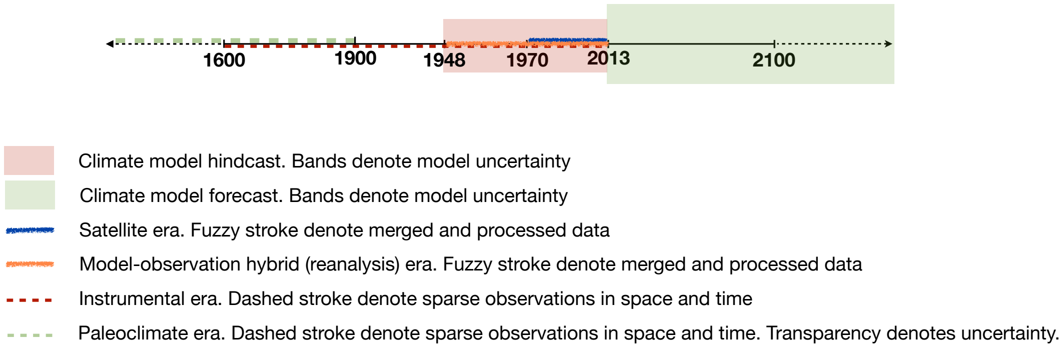

1600
1900
1948
1970
2013
2100
Climate model hindcast. Bands denote model uncertainty
Climate model forecast. Bands denote model uncertainty
Satellite era. Fuzzy stroke denote merged and processed data
Model-observation hybrid (reanalysis) era. Fuzzy stroke denote merged and processed data
Instrumental era. Dashed stroke denote sparse observations in space and time
Paleoclimate era. Dashed stroke denote sparse observations in space and time. Transparency denotes uncertainty.

Figure 2.1: Climate science has numerous types of data, each with its own challenges.

The majority of climate data available can be classified into four categories based on their source: in-situ, remote sensed, model output, and paleoclimatic.

In-situ records of climate data date back to the mid- to late 1600s [189]. Today, observational data are gathered from a plethora of in-situ instruments such as ships, buoys, and weather balloons. Such data tend to be sparse measurements in space and time since they are only available when measurements are gathered and where the instrument is physically located. For example, a weather balloon records frequent measurement only for a limited time duration and at its physical location. Additionally, raw measurements can be noisy due to measurement error or other phenomena temporarily impacting measurement (*e.g.* strong winds affecting temperature measurements). A final caveat is such data are dependent on the geopolitical state of where the instruments are deployed. For instance, the quality of sea surface temperatures along the Atlantic ocean decreased during World War II due to reduced reconnaissance.

Remote sensed satellite data became available in the late 1960s and are a great source of relatively high quality data for large portions of the earth. Although they are considered one of the best sources of global observational data, remote sensed satellite data have notable limitations. First, satellite data are subject to measurement noise and missing data due to obstructions from clouds or changes in orbit. Second, due to their short life-span ($\sim$ a decade) and evolving technology, satellite data can be heterogeneous.

Currently, the biggest contributors to climate data volume are climate model simulations. Climate models are used to simulate future climate change under various scenarios as well as reconstructing past climate (hindcasts). Such models run solely based on the thermodynamics and physics that govern the atmosphere-hydrosphere-land surface system, with observational data used for initialization. While these data tend to be spatio-temporally continuous, they are highly variable due to the output's dependence on parameterization and initial conditions. Furthermore, all model output come with inherent uncertainties given that not all the physics are resolved within models and our incomplete understanding of many physical processes. Therefore, the climate science community often relies on multi-model ensembles where numerous model outputs using various parameters and initial conditions are averaged to mitigate the uncertainty any single model output might have. For instance, the Nobel Peace Prize winning Intergovernmental Panel on Climate Change (IPCC) used multi-model ensembles to present its assessment of future climate change [231]. Finally, there still exist several theoretical and computational limitations that cause climate models to poorly simulate certain phenomena, such as precipitation.

To address the noisy and heterogeneous quality of in-situ and satellite observations, a new generation of simulation-observation hybrid data (or reanalyses) have emerged.

Reanalysis datasets are assimilated remote and in-situ sensor measurements through a numerical climate model. Reanalyses are generated through an unchanging (”frozen”) data assimilation scheme and models that take all available observation from in-situ and remote sensed data every 6-12 hours over a pre-defined period being analyzed (*e.g.* 1948–2013) [1]. This unchanging framework provides a dynamically consistent estimate of the climate state at each time step. As a result, reanalysis datasets tend to be more physically consistent (*i.e.* obey autocorrelation and energy conservation laws, *etc.*) than the raw observational records and have extended spatio-temporal coverage. While reanalyses are considered the best available proxy for global observations, their quality is still dependent on that of the observations, the (assimilation) model used, and processing methods. More domain specific quality issues for certain reanalysis data can be found at `http://www.ecmwf.int/research/era/do/get/index/QualityIssues`.

Finally, researchers have been reconstructing historical data using paleoclimatic proxy records such as trees, dunes, shells, oxygen isotope content and other sediments [2]. Such data are used to study climate variability at the centennial and millennial scales. Given the relatively short record of observational data, paleoclimate data are crucial for understanding pre-instrumental climate variability. It is important to note that paleoclimate data are proxies, such as using tree rings to infer rainfall trends. Furthermore, such records are used to infer climate over a wide time-span and the time of occurrence cannot be exact. Finally, paleoclimate techniques are still developing and quality testing methods continue to be an active area of research.

2.0.2 Unique Characteristics of Climate Data

In the introduction, we briefly mentioned some of the data's characteristics and in the previous subsection we discussed some of the issues that surround data quality and availability. In this section, we expand further on this subject to provide the reader with a more nuanced discussion of climate data characteristics.

From a modeling perspective, the most fundamental difference between traditional (categorical) and spatio-temporal climate data is that data that are close in space and time tend to be more similar than data far apart. This “first law of geography” which

[1] `http://climatedataguide.ucar.edu/reanalysis/atmospheric-reanalysis-overview-comparison-tables`

[2] `http://www.ncdc.noaa.gov/paleoclimate-data`

is more commonly known as *autocorrelation* dictates that spatio-temporal data not be modeled as statistically independent [233]. As a result, models that assume independent and identically distributed (*i.i.d*) observations will be limited in modeling climate data and their underlying processes.

Another notable difference is that spatio-temporal phenomena in climate are not concrete "objects" but evolving patterns over space and time. For example, a hurricane doesn't simply appear and disappear, rather an atmospheric instability slowly evolves into a hurricane that gradually gains strength, plateaus, and gradually dissipates over a spatio-temporal span. This is a profound difference from traditional (binary) data mining where objects are either present or absent. Such spatio-temporal evolutionary processes are well captured by the differential equations used in climate models. While differential equations are costly to solve and have other well-known limitations, data mining has no (cost efficient) statistical analog to model the evolutionary nature of spatio-temporal phenomena [62]. This is becoming a significant challenge and efforts are emerging, especially within the spatio-temporal statistics community, to provide an alternative. However such methods have yet to gain wide applicability.

Another fundamental difference in climate data is the uncertainty, variability, and diversity inherent in such datasets. Uncertainty in climate data stems from the fact that many climate datasets have biases in sampling and measurement, along with some datasets being the product of merged (uncertain) data. Furthermore, researchers are seldom provided with the data's uncertainty information. For instance, there are datasets that span the past 150 years, and while it is reasonable to assume that older data are less reliable, often there is no way to objectively characterize such uncertainty. Alternatively, if one chooses to restrict their attention to the most reliable data periods (post 1979), then a data-driven research agenda becomes more challenging due to the short record.

Climate data tend to also be highly variable. Sources of variability include: (i) natural variability, where wide-range fluctuations within a single field exist between different locations on the globe, as well as at the same location across time; (ii) variability from measurement errors; (iii) variability from model parameterization; and (iv) variability from our limited understanding of how the world functions (*i.e.* model representation). Even if one accounts for such variability, it is not clear if these biases are additive and there are limited approaches to de-convolute such biases a posteriori.

We refer to data diversity as its heterogeneity in space and time. That is data are available at various spatio-temporal resolutions, from different sources, and for different uses. Often times, a researcher must rely on multiple sources of information and adequately integrating such diverse data remains a challenges. For example, one may have access to three different sea surface temperature datasets: one reanalysis dataset at a 2.5° resolution, another reanalysis dataset at 0.75° resolution, and a satellite dataset at 0.25° resolution. It is unclear how

Climate phenomena operate and interact on multiple spatio-temporal scales. For example changes in global atmospheric circulation patterns may have significant impacts on local infrastructures that cannot be unearthed if studying climate only at a global scale (*i.e* "will global warming cause a more rainy winter in California in year 2020?"). Understanding such multi-scale dependencies and interactions is of significant societal interest as there is a need to provide meaningful risk assessments about global climate's impact on local communities.

Finally, many climate phenomena have effects that are delayed in space and time. Although "long-range" relationships do exist in traditional data mining applications, such as a purchase occurring due to a distant acquaintance recommending a product, they are far more complex in a climate setting. Relationships in climate datasets can not only be long-range in both space and time as well as multivariate, there are exponentially many space-time-variable subsets where relationships may exist. As a result, identifying significant spatio-temporal patterns depends on knowing what to search for as much as *where* to search for such a pattern (*i.e.* which spatio-temporal resolution).

In the next chapter, we will provide the reader with a concise review of the STDM literature pertaining to climate data.

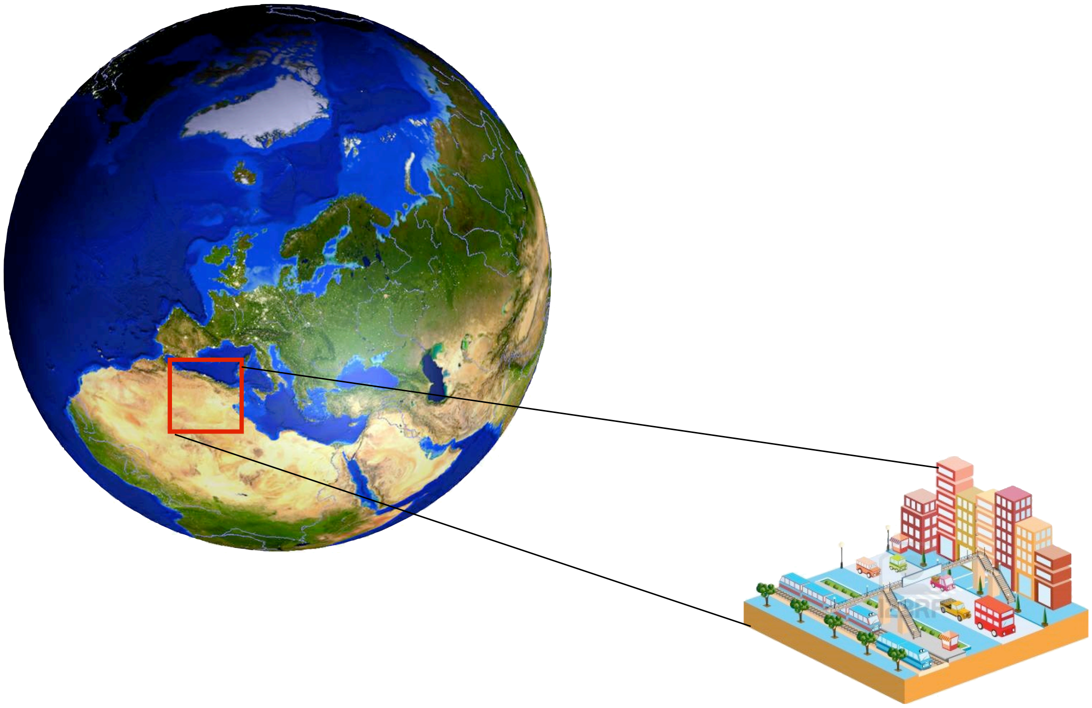


Figure 2.2: A large amount of climate data is at at global spatial scale ($\sim$250km), however many climate-related questions are at the regional ($\sim$50km) or local (km or sum-km) scale. This multi-scale discrepancy is a significant data mining challenge.

Chapter 3

Advances in STDM applications to Climate

Although the fields of temporal and spatial data mining research are relatively mature [215, 167], STDM is an emerging computer science field. The main driver for such emergence is the growth in spatio-temporal datasets and associated real-world challenges. Broadly speaking, STDM originated in the form of extending temporal capabilities to spatial data mining problems, or accommodating for space in temporal data mining applications. The former extension is a rather natural one given the widespread availability of time-stamped geographic data. Intuitively, one may think of the spatio-temporal context of the data as *constraints* for a knowledge discovery algorithm. Expert constraints have been a staple of knowledge discovery algorithms as they have the potential to improve a model's scalability (by reducing the search space), accuracy (by discarding implausible models) and interpretability [57, 56, 168, 71]. In the same spirit, one may think of spatio-temporal information as expert constrains on traditional learning algorithms. However, a constraint point-of-view cannot be adopted for many existing algorithms given the strong assumptions such methods have on the nature of the data (*e.g.* i.i.d) or the data generation process (Gaussian, Poisson, *etc.*) In this case, an entire new generation of learning algorithms must be developed to account for the specific nature of *data* and *problems* STDM is trying to address. In this section, we expose the reader to a broad range of STDM application to climate. In the following sections, we will provide a simple introduction and example for each broad type of applications as well as a sample of the literature within those applications.

3.1 Spatio-Temporal Query Matching

Some of the earliest works in STDM were in the context of earth and climate sciences. Intuitively, the first step a data miner undertakes is exploring the data and its characteristics. Given the large size of climate data, early priorities were focused on data exploration and collaborative analysis.

Mesrobian et al. [175] introduced CONQUEST, a parallel query processing system for exploratory research using geoscience data. The tool allowed scientists to formulate and mine queries in large datasets. This is one of the first works to track distortions in a continuous field. One application demonstrated in their work was the tracking of cyclones as local minima within a closed contour sea level pressure (SLP) field [175, 226]. As en extension to CONQUEST, Stolorz and Dean [225] introduced Quakefinder, an automatic application that detects and measures tectonic activity from remote sensing imagery. Mesrobian et al. [176] introduced Oasis a extensible exploratory data mining tool for geophysical data. A similar application is the algorithm development and mining framework (ADaM) [200] which was developed to mine geophysical events in spatio-temporal data. Finally, Baldocchi et al. [12] introduced FLUXNET a collaborative research tool to study the spatial and temporal variability of carbon dioxide, water vapor, and energy flux densities.

The early emphasis of all these works was on scalable query matching as well as abstracting the data and their formats to the researcher to focus more on exploratory research rather than data management. However, large-scale collaborative research efforts are costly and require extensive infrastructures and management, effectively increasing the risk associated with such endeavors. Furthermore, we often embark on exploratory research without prior knowledge of the patterns of interest making explicit query searches non-trivial. Finally, such exploratory efforts should capitalize on the recent advances in both spatial and temporal subsequence pattern mining (*e.g.* [96, 199]).

3.2 Pattern Mining

One of the fundamental applications of data mining is finding patterns within a dataset. Pattern mining refers to the insightful grouping of features that share similar characteristics such as statistical properties or frequency of occurrence. In this section we will review three notable pattern mining approaches within climate applications: empirical orthogonal function (EOF) analysis, clustering, and user-defined pattern mining.

One of the most fundamental tools in spatio-temporal pattern finding is empirical orthogonal function (EOF) analysis. EOFs are synonymous to the eigenvectors in traditional eigenvalue decomposition of a covariance matrix. As pointed out by Cressie and Wikle [62], in the discrete case, EOF analysis is simply principle component analysis (PCA). In the continuous case, it is a Karhunen-Loève (K-L) expansion. EOF analysis has been traditionally used to identify a low dimensional subspace that best explains the data's spatio-temporal variance. By taking the data's first principal component, researchers seek to identify dominant spatial structures and their evolution over time. For instance, Mestas-Nuñez and Enfield [177] analyzed the rotated[1] EOFs of global SST data and linked the first six principal components to ocean-atmospheric modes[2]. In another application, Basak et al. [16] used independent component analysis to discover the North Atlantic Oscillation index (NAO) [161] in SLP data. For a comprehensive discussion of EOF analysis for climate data please see [257].

Within clustering applications, Hoffman et al. [130] developed a spatio-temporal clustering algorithm to identify regions with similar environmental characteristics. White et al. [256] applied the techniques presented in [130] to generate climate and vegetation clusters that were subsequently used to infer phenological responses to climate change. Braverman and Fetzer [27] mined large-scale structures in climate data using a data compression technique based on entropy-constrained vector quantization [52] to generate multivariate distribution estimates of the data and monitored the changes of such distributions across space, time, and resolution. McGuire et al. [174] used spatial neighborhood and temporal discretization methods to identify spatio-temporal neighborhoods in SST data. In another clustering application, Gaffney et al. [107] clustered cyclone tracks using a regression mixture model and works by Camargo et al. [29] and

[1] rotation transforms the EOF into a non-orthogonal linear basis

[2] Emanuel [88] points out that EOFs are *not* mathematically equivalent to modes

Camargo et al. [30] further analyzed the clusters to discuss various properties of tracks belonging to each cluster. Although there are numerous works in the field, finding significant spatio-temporal clusters remains a major challenge because of both spatial and temporal variability. In particular, the physical meaning and significance of clusters are sometimes debatable. Furthermore, traditional feature similarity measures used to assign features to clusters, such as Euclidean distance from cluster centroids, might not have a physical meaning in climate applications.

Finally, sample works that mined climate data for user-defined patterns include: automatically identifying and tracking cyclones in the atmosphere as close contoured negative anomalies in SLP data. There are several techniques to find and monitor such patterns as storm monitoring is an active field of research. For a review please see [240]. Another dominant climate pattern is the InterTropical Convergence Zone (ITCZ) a phenomena on a daily time scale over the east Pacific. Bain et al. [11] developed a spatio-temporal Markov random field to detect the ITCZ in satellite data. Henke et al. [126] extended such methods by using a super- and semi-supervised method to track this dynamic phenomena and its properties in satellite and infrared data. Within pattern finding applications, a large number of climate phenomena tend to exist within specific spatio-temporal subsets. Naively searching for such subsets is prone to combinatorial explosion due to the exponentially-many subsets in both space and time. A notable emerging pattern mining application is that of identifying user-defined patterns in large data. Figure 7.2 shows an example of pattern mining in continuous spatio-temporal climate data. Ocean eddies (rotating whirlpools in the ocean) manifest in numerous climate datasets and extracting such a pattern from noisy climate data is an active field of research. In this case, the pattern of interest is localized sea surface height anomalies spanning 50 to 100s of kilometers over time-spans of weeks to months. The goal is to identify such patterns on a global scale. We will discuss this application in depth in the next section.

3.3 Event and Anomaly Detection

Automatic identification of climate events such as global changes in vegetation, droughts, and extreme rainfall is of interest to a variety of researchers. In climate applications,

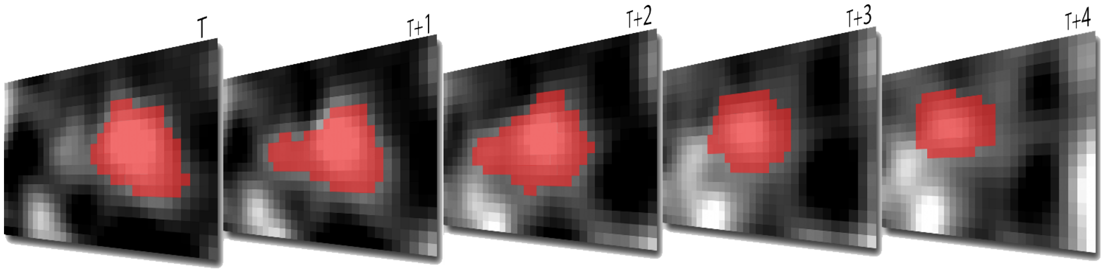


Figure 3.1: An ocean eddy moving in time as detected in ocean data. One of the challenges of STDM is to identify significant patterns in continuous spatio-temporal climate data.

an event is an instance in time when a significant and persistent change occurs. In contrast, an anomaly (or outlier) is a short yet significant deviation from normal behavior. Figure 3.2 shows examples for an event and an anomaly. The time-series denote changes in vegetation over time as defined by remote sensed images. Panel (a) shows relatively stable vegetation from 2000 until 2003 when a distinctly new and persistent vegetation pattern emerged. Mid-2003 would be considered an event change point, where the vegetation level significantly and persistently changed from the previous period. Panel (b) shows a sudden drop in vegetation due to a forest fire in 2006. The vegetation level did recover after a few years. As a result the fire event can be considered an anomaly.

A number of studies have monitored event and anomaly changes in ecosystems data. Boriah et al. [26] proposed a recursive merging algorithm that exploited the data's seasonality to distinguish between locations that experienced a land cover change and those that did not. Mithal et al. [179] introduced a global land-cover change algorithm that accounted for the natural variability of vegetation levels. While the land-cover change literature is vast, especially within the remote sensing community, Mithal et al. [178] provide a concise discussion of STDM techniques and challenges related to land-cover change. In another global-scale event detection application, Fu et al. [106] extended the traditional Markov random field (MRF) model [249] used in spatial statistics by maintaining the spatio-temporal dependency structure of the MRF to autonomously detect droughts globally.

There is extensive STDM work for outlier detection for disease outbreaks [183, 182]

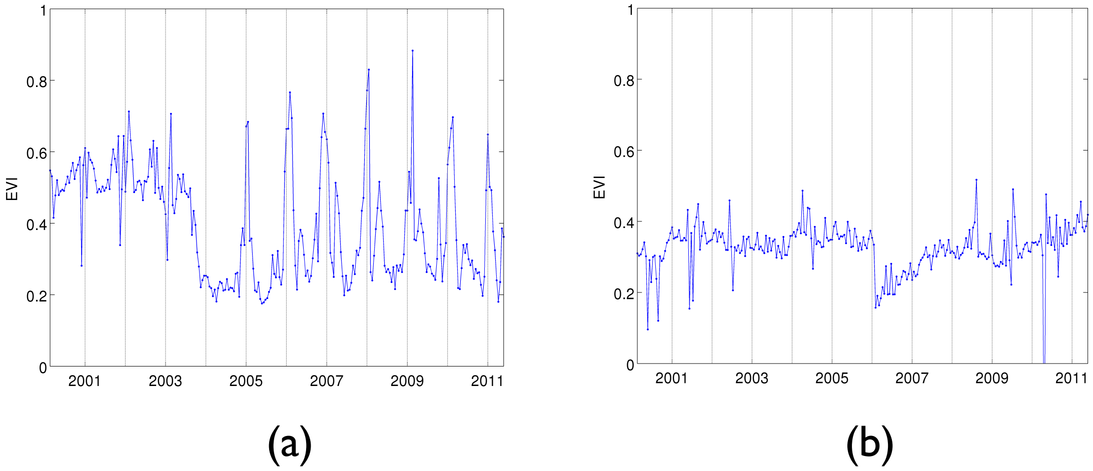


Figure 3.2: An example of a spatio-temporal event (a) and anomaly (b). The time-series denote changes in vegetation over time. (a) A land-cover change event as seen in the decrease of vegetation due to agricultural expansion in 2003. (b) an abrupt drop in vegetation due to a forest fire in 2006, the vegetation gradually returned after the fire.

and the climate applications base their work on that domain. To address the fact that atmospheric events occur at different scale in space and time, Cheng and Li [51] developed a multi-scale spatio-temporal outlier detection algorithm by evaluating the change between consecutive spatial and temporal scales to detect abnormal coastal changes. Barua and Alhajj [15] used a parallel wavelet transform to detect spatio-temporal outliers in SST data. Wu et al. [260, 261] detected spatio-temporal outliers in precipitation data by storing high discrepancy spatial regions over time in a tree. The authors were able to recover anomalous precipitation spatio-temporal spans that closely mimic the El-Niño Southern Oscillation cycle. Anbaroğlu [6] used a space-time autoregressive integrated moving average to define coherent spatio-temporal neighborhoods. An outlier was then defined if its value was significantly different from the mean that of nearby spatio-temporal neighborhoods.

Although traditional data mining has extensive research on event and outlier detection [41], there are notable differences that make such applications within climate extremely challenging. First, unlike traditional data mining where events are relatively unambiguous (*e.g.* a purchase, check in, *etc.*) the very pattern that represents an event is not known in advance or might vary based on a spatio-temporal context (*e.g.*

different precipitation events could be labeled as a flood or drought depending on the time and location of occurrence). Second, climate data tends to be noisy and highly variable therefore one cannot simply label anomalous events as a large deviation from the mean. For instance, Ghosh et al. [109] used an extreme value theory method to highlight the fact that due to high spatial variability, anomaly detection must be in relation to space and time. Third, it is challenging to distinguish a measurement error (*i.e.* a spurious anomaly) from a low-probability event. Sugihara and May [228] proposed a method to distinguish between chaos and measurement error using short-term predictability, however additional advances might be needed. Finally, there is extreme societal interest in identifying prolonged dramatic changes in climate, known as climate state shifts [209]. Such events are critical because species tend to be less resilient to such severe abrupt changes (*e.g.* a region suddenly transforming into a desert). However, given the relatively small number of years with high quality data, it is difficult to establish with certainty whether an observed change is a significant shift or a mere fluctuation if taken into the proper spatio-temporal context. Therefore there is a need to develop novel event significance tests that would account for the limited number of reliable observations within certain datasets.

3.4 Relationship Mining

Within climate applications, researchers are interested in linking changes in one variables (e.g. global temperatures) to other phenomena (*e.g.* land cover or total number of hurricanes). A common example is relating changes in Pacific sea surface temperatures (SST), known as El-Niño Southern Oscillation (ENSO), to other global phenomena. To abstract the complex ENSO phenomenon, researchers use the mean SST of fixed regions in the Pacific to construct NINO indices and subsequently relate them to other phenomena. Figure 3.3 shows the linear correlation coefficients between one such NINO indices (NINO1+2) and global land surface temperature anomalies. The figure suggests that when the NINO1+2 is in a positive extreme, land temperatures tend to be high in South America, while land temperatures tend to be cooler in the south eastern United States. There are numerous works that analyze linear relationships between climate variables. Goldenberg and Shapiro [111] used linear and partial linear correlations to

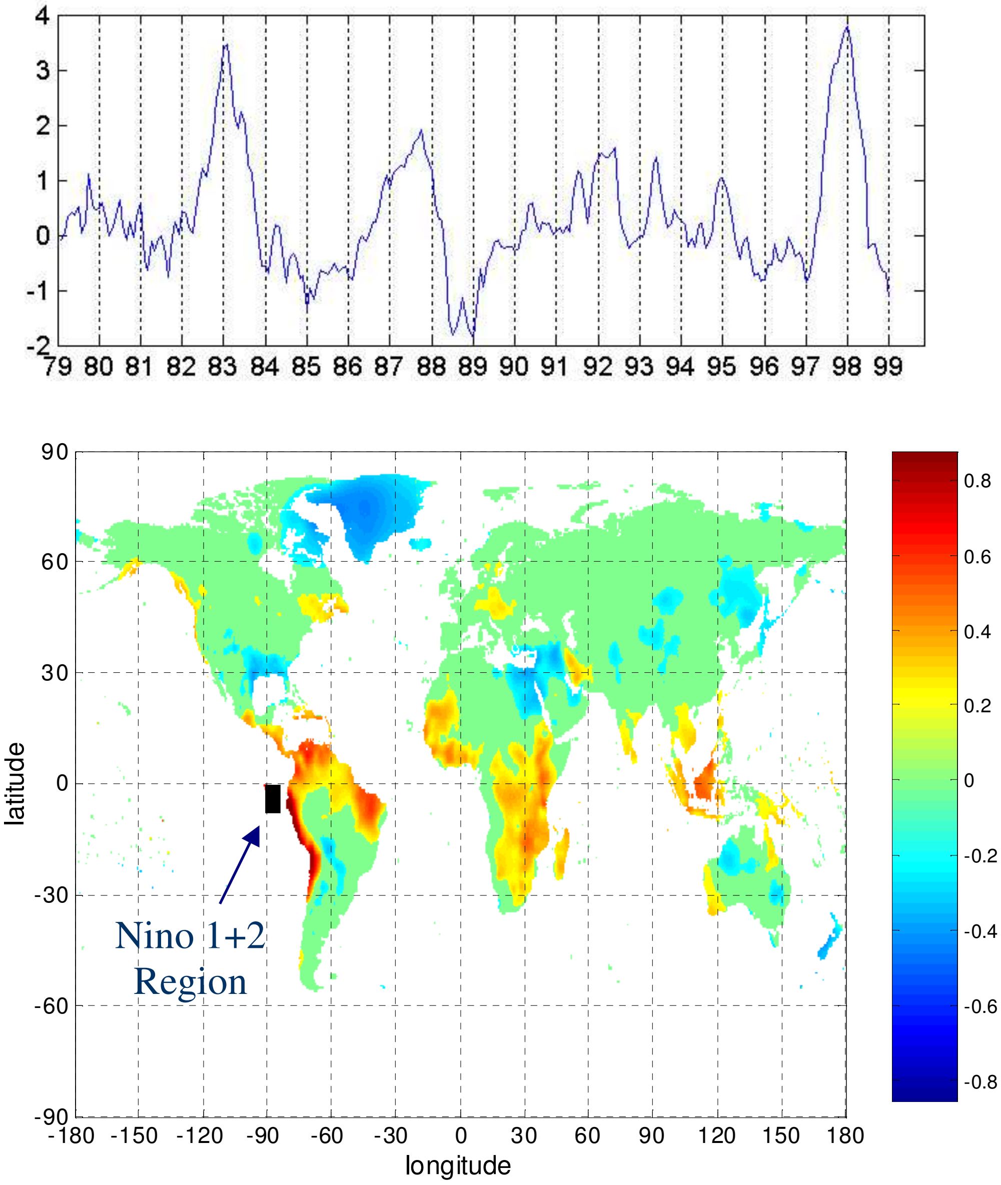


Figure 3.3: Top: The NINO1+2 time-series which was constructed by averaging the sea surface temperatures (SST) of the box highlighted in the map below. Bottom: the linear correlation between the NINO1+2 index and global land surface temperature anomalies.

link vertical wind shear in the Atlantic to SST and Sahel rainfall patterns. Webster et al. [255] analyzed the linear correlation between basin-wide mean SST and seasonal TC counts in all the major basins between 1970-2005 and concluded that the upward trend in Atlantic TC seasonal counts cannot be attributed to the increased SST. This was because not all basins that had an increase in SST, had a corresponding increase in TC counts. In another study, Chen et al. [50] used the sea surface temperatures and found different oceanic regions correlate with fire activity in different parts of Amazon. There are numerous other studies like the ones mentioned above, however detecting relationships in large climate datasets remains extremely challenging. For example, the data used in [50] only spanned 10 years (N=10). It is also impossible to isolate all confounding factors in global climate studies since many conditions can affect any given phenomenon.

One other limitation of linear correlation is its inability to capture nonlinear relationships. While there are studies that use nonlinear measures such as mutual information (*e.g.* [136]), climate scientist use *composite analysis* as a another way to quantify how well one variable explains another. Figure 3.4 shows an example of how composites are constructed. For a given anomaly index, in this case NINO3.4 index, we can identify extreme years as those that significantly deviate from the long-term mean (*e.g.* less/greater than one or two standard deviations). The time-series in Figure 3.4's upper panel highlights the extreme positive (red squares) and negative (blue squares) years within the NINO3.4 index from 1979 to 2010. Using the extreme positive and negative years, one can comment on how a variable responds to the extreme phases of a variable (in this case the NINO3.4 index). Take the June-October mean vertical wind shear over the Atlantic basin (Figure 3.4 bottom panel). The composite shows the difference between the mean June-October vertical wind shear during the 5 negative extreme years and the 5 positive extreme years. The bottom panel suggests that extreme negative years in NINO3.4 tend to have low vertical wind shear along the tropical Atlantic. One of the advantages of using composite analysis is that it does not make specific assumptions about the relationship between the two variables, it could be linear or non-linear. One must also use caution when analyzing composites. While we can test the significance in the difference in means between the positive and negative years, traditional significance tests assume independent observations which might not be the case for such data. Furthermore, the sample size of extreme events might be too small to be significant. For example, Kim and Han [146] constructed composites of Atlantic hurricane tracks based on the warming patterns in the Pacific ocean. One phase of their index had a sample size of 5 years (out of 39 years). To test the significance of the composite that summarized hurricane tracks during those years, the authors used a bootstrapping technique [75] to determine how significant was the mean of the small sample relative to random noise.

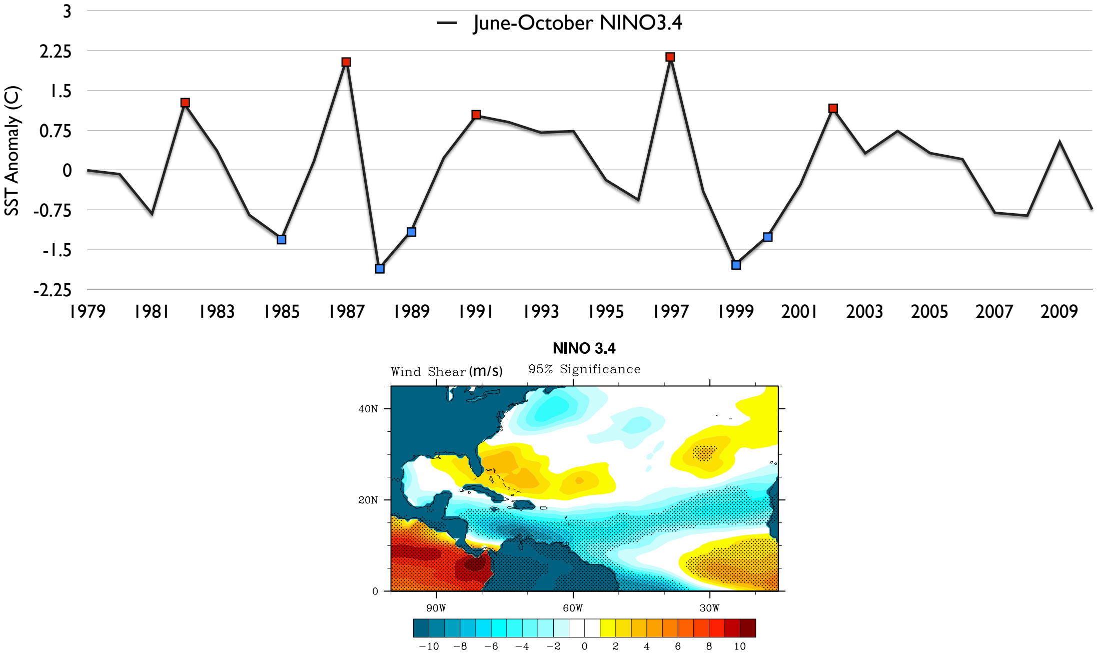


Figure 3.4: An example on how composites un-earth non-linear relationships between variables. Top panel: time-series of SST anomalies in the NINO3.4 region. Bottom panel: Composite of June-October mean vertical wind shear, which was constructed by subtracting the top panel's mean of the negative extremes from the mean of the positive extremes. The figure shows that warming in the Pacific ocean has significant impact on an other variable in the tropical Atlantic.

Finally, given that one searches for potential relationships (linear or non-linear) between a large number of observations, the likelihood of observing a strong relationship by random chance is higher than normal (known as multiple hypothesis testing or field significance). Figure 3.5 shows an example of the same dataset (geopotential height) correlated with a real index (left) and random noise (right). The figure shows how easily a random pattern can yield misleadingly high correlations with smooth spatial patterns.

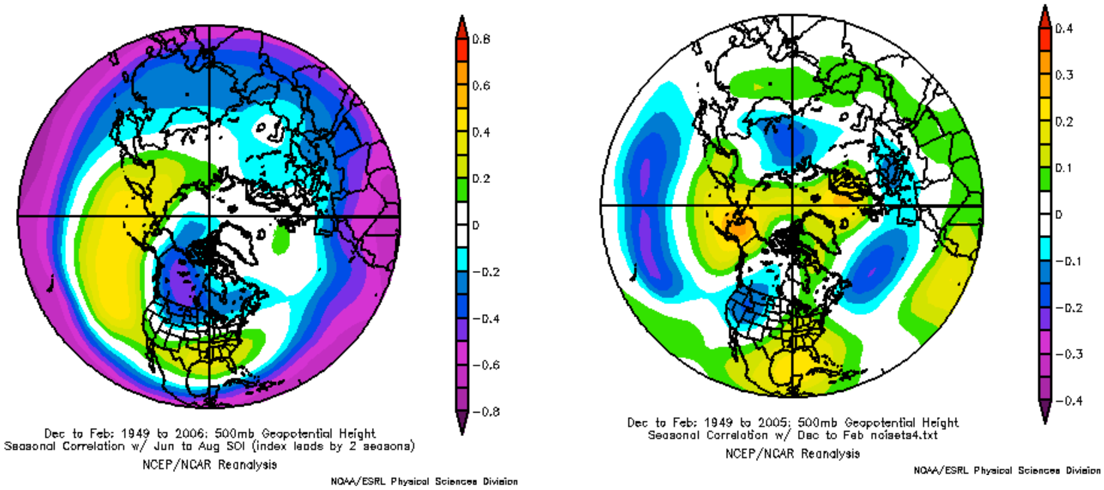

0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
Dec to Feb: 1949 to 2006: 500mb Geopotential Height
Seasonal Correlation w/ Jun to Aug SOI (index leads by 2 seasons)
NCEP/NCAR Reanalysis
NOAA/ESRL Physical Sciences Division
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
Dec to Feb: 1949 to 2005: 500mb Geopotential Height
Seasonal Correlation w/ Dec to Feb noisets4.txt
NCEP/NCAR Reanalysis
NOAA/ESRL Physical Sciences Division

Figure 3.5: Geopotential height correlated with the Southern Oscillation Index (SOI; left) and random noise (right). This is an example how high and spatially coherent correlations can be the result of random chance.

3.5 Spatio-Temporal Predictive Modeling

One of the major applications to climate is the ability to model and subsequently predict future phenomena. Statistical models hold great promise to model phenomena not well resolved in physics based models, such as precipitation. With the growth of statistical machine learning there have been numerous works on predictive modeling. In this section, we will mainly focus on some of the works that explicitly addressed the spatio-temporal nature of the data.

Coe and Stern [58] used a first- and second-order Markov chain to model precipitation. However scarce observations at the time almost certainly limit the generalization of such an approach. Cox and Isham [60] proposed a spatio-temporal model of rainfall where storm cells obey a Poisson process in space and time with each cell moving at random velocity and for a random duration. Additional reviews of precipitation models can be found in [259, 218, 221]. Huang and Cressie [137] improved on traditional spatial prediction models of water content in snow cover (also known as snow water equivalent) using a Kalman filter-based spatio-temporal model. The model effectively incorporated snow content from previous dates to make accurate snow water equivalent predictions for locations where such data was missing. Cressie et al. [63] designed a spatio-temporal prediction model to model precipitation over North America. Their work employed random sets to leverage data from multiple model realizations (*i.e.* multiple initial conditions, parameter settings *etc.*) of a North American regional climate model.

van Leeuwen et al. [241] built a logistic regression-based model trained on land surface temperatures to detect changes in tropical forest cover. Karpatne et al. [143] extended the work in [241] by addressing the heterogeneous nature land cover data. Instead of training a single global model of land cover change based on a single variable (*e.g.* land surface temperature), they built multiple models based on land cover type to improve single-variable forest cover estimation models. A related application within the field of land cover change is autonomously identifying the different types of land-cover (urban, grass, corn, *etc.*) based on the pixel intensity of a remote sensed image. Traditional remote sensing techniques train a classifier to classify each pixel in an image to belong to certain land-cover class [230]. However, each pixel is classified independently of every other pixel without any regard for the spatio-temporal context. This causes highly variable class labels for the same pixel across time. Mithal et al. [180] improve the classification accuracy of existing models by considering the temporal evolution of the class labels of each pixel.

One of the major challenges in predictive modeling is that climate phenomena tend to have spatial and temporal lags where distant events in space and time affect seemingly unrelated phenomena far away (physically and temporally). Therefore identifying meaningful predictors in the proper spatio-temporal range is difficult. It is also important to note that certain extreme events that are of interest to the community (*e.g.* hurricanes) are so rare that the number of observations is much smaller than the data's dimensionality ($n << D$). In this case, a minimum number of predictors must be used to avoid overfitting and a poor generalized performance. For instance, Chatterjee et al. [43] used a sparse regularized regression method to identify the interplay between oceanic and land variables in several regions around the globe (*e.g.* how does warming in the South Atlantic affect rainfall in Brazil?). Their use of parsimony significantly improved the model's performance. Finally, model interpretability is crucial for spatio-temporal predictive modeling because the majority of climate science applications need a physical explanation to be adopted by climate scientists.

3.6 Network-based Analysis

For gridded climate data, numerous efforts have sought to abstract the large complex data and associated interactions into a simple network. Generally, nodes in the climate network are geographical locations on the grid and the edge weights measure a degree of similarity between the behavior of the time-series that characterize each node (*e.g.* linear correlation [237], mutual information [73], syntonization [8], *etc.*) Once a network is built, it is possible to apply the techniques previously discussed such as relationship mining [144], predictive modeling [224, 211], or pattern mining [223] on the transformed data.

Steinbach et al. [223] were one of the first to organize climate data into a network and applied a shared nearest neighbor algorithm on the network to discover the strongest climate indices: time-series that abstract the state of the atmosphere over large spatial and temporal spans. Kawale et al. [144] extended the work in [223] to allow for dynamic dipoles (strongly correlated distant spatial regions) in climate data. Kawale et al. [145] proposed a bootstrapping method to test the significance of such long-range spatio-temporal patterns.

Inspired by complex networks, [237] were the first to propose the notion of a *climate network* and analyze its properties and how they relate to physical phenomena. For example, several studies have found the network structure to correlate with the dominant large-scale signals of global climate such as El-Niño [74, 263, 121]. Similarly, Tsonis et al. [238] showed that some climate phenomena and datasets obey a small-world network property [251]. Furthermore, several studies found distinct structural differences between the networks around tropical and extra-tropical regions [238, 73]. Berezin et al. [22] analyzed the evolution and stability of such networks over time and found that networks along the tropics tend to be more stable. Other studies have linked regions with high in-bound edges, known as supernodes, to be associated with major large-scale climate phenomena such as the North Atlantic Oscillation [238, 239].

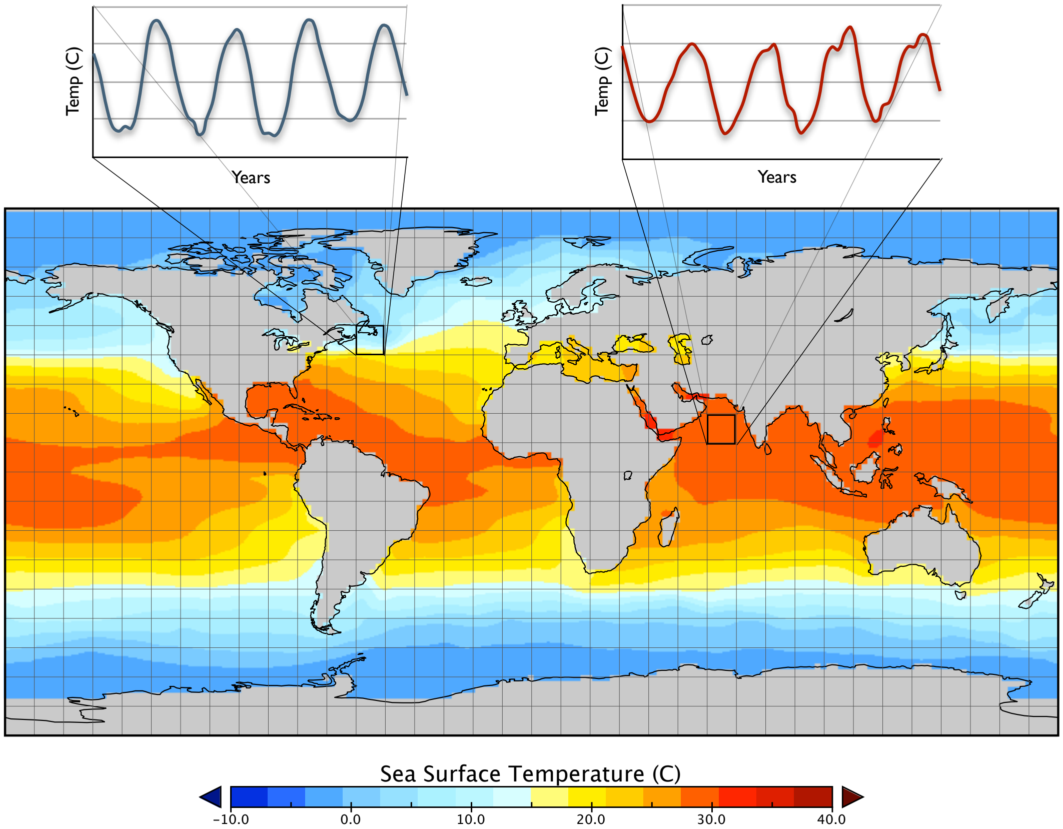


Figure 3.6: Gridded spatio-temporal climate data can be analyzed in a network format. Each grid location is characterized by a time series. A network can be constructed between each location with an edge weight being the relationship between the time-series of each location.

Others have built networks using non-gridded discrete climate data. Elsner et al. [82] used seasonal hurricane time-series to construct a network to study interannual hurricane count variability. Fogarty et al. [102] built a network to analyze coastal locations (nodes) and their associated hurricane activity (edges) and found distinct connectivity difference between active and inactive regions. Furthermore, the authors connected various network topographies to phases of the El-Niño Southern Oscillation.

While network-based methods within climate are increasingly popular, these efforts are relatively young and several questions remain such as how to sparcify fully connected networks, the notion of multi-variate climate networks, and the distinction between statistical and physical connectivity [192].

We will spend the remainder of the chapter demonstrating a case study of spatio-temporal pattern mining with an autonomous ocean eddy monitoring application. This is because ocean eddies are a central part of ocean dynamics and impact marine and terrestrial ecosystems. Furthermore, identifying and tracking eddies form a new generation of data mining challenges where we are interested in tracking uncertain features in a continuous field.

Chapter 4

Spatio-Temporal Data Mining and Tropical Cyclones

4.1 Background

Global climate change and its effects on Atlantic tropical cyclone (TC) activity has become one of the most contested issues in climate science [171, 236, 3]. The difficulty of attributing a change in TC frequency to global climate change stems from the lack of reliable historical data [42, 165], the large amplitude fluctuations in present-day storms [70], and knowledge gaps converning the exact influence various climate factors have on TC activity [119, 88]. This last issue is the focus of the current chapter. Although the relationship between TCs and various climate factors such as sea surface temperature (SST) have been posited [119, 208], it is unclear how warmer SST will interact with and impact other factors that influence TC activity. Given the observed upward trend in Atlantic SST in recent decades [207] (Figure 4.1), understanding the relationship between SST and future TC activity is crucial.

Currently, a clear understanding of TC formation (cyclogenesis) in a warming environment is still lacking [88, 195], effectively making TC frequency predictions highly uncertain. Theoretical limitations are highlighted in the current high-resolution models' failure to consistently predict an increase or decrease in the total number of TCs in a warming environment. The majority of global circulation models (GCMs) forecast a decrease in the total number of TCs as the atmosphere continues to warm. Locally,

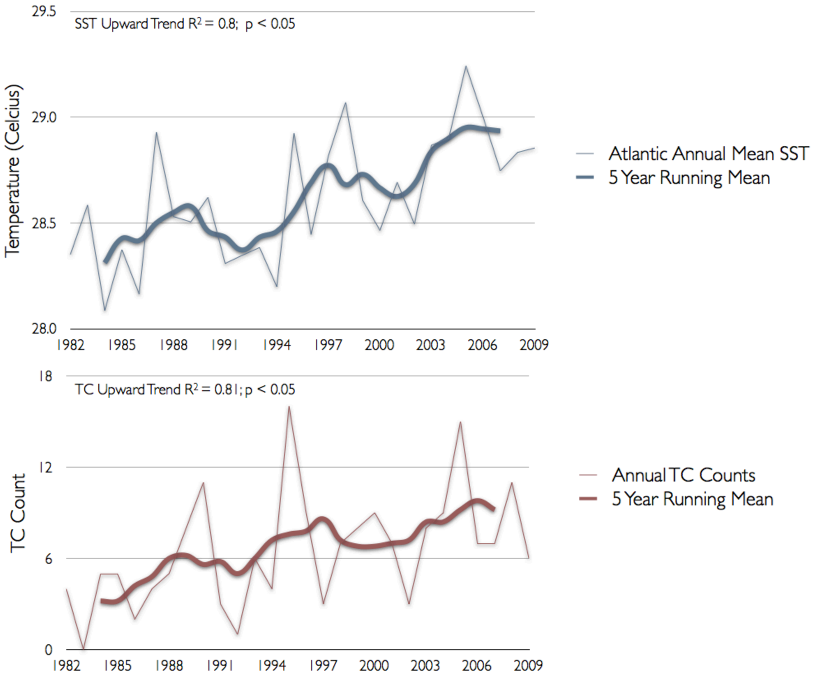


Figure 4.1: **Upward trend in Atlantic basin SST and TC frequency**. *Top:* Mean annual SST (June-October) averaged over the Atlantic basin (5° to 30° N and 20° to 90°W). *Bottom:* Annual TC counts (June-October) for the Atlantic basin. Both datasets show an upward trend. Although given the short length of the storm counts time-series, there is considerable debate whether there is an upward TC trend or natural variability [255, 39, 254].

however, regional circulation models (RCMs) have been significantly more uncertain with projected changes of up to $+/-50\%$ from current frequency [155].

In addition to the catastrophic physical and financial aftermaths, a clearer understanding of future TC activity is crucial for several reasons. First, it is believed that TCs play a role in the the ocean's poleward heat transport. TCs' high wind speeds and strong internal waves cause vertical mixing between the ocean's (warm) near surface mixed layer and the (cool) main thermocline base. This mixing causes the mixed-layer to cool and warms the thermocline [197]. Subsequently, mixing induces a poleward heat transport and causes a net increase in ocean heat content [85, 222]. The heat transport has three implications: First the injection of heat at higher latitudes might mitigate the warming in the tropics by amplifying the warming of higher latitudes. Second, the warming of higher latitudes creates a slower pole-to-tropics heat gradient, which can affect other phenomena. Finally, this transport could explain some the increased sea surface temperature (SST) in the northern Atlantic that has been attributed to global warming [235, 231] as well as reduced sea-ice in the north Atlantic and Arctic oceans [59].

Second, while a majority of TC research is concerned with landfall prediction, TCs that form and dissipate over the ocean are also important to study. The previously described mixing has also been shown to enhance ocean primary (phytoplankton) production. Primary production has a significant impact on marine ecosystems as the base of the ocean's food chain and the ocean's major oxygen producer. More importantly, phytoplankton affects the uptake of carbon dioxide, a greenhouse gas linked to natural and anthropogenic climate change [91, 169]. The previously described vertical mixing enhances phytoplankton production by raising the deep-layer nutrient-rich water, especially near the core of the storm [169, 123]. However, it is possible that storms might also delay phytoplankton bloom by up to two weeks for those outside the storm's core [123]. Additionally, ocean mixing and transport causes the phytoplankton to migrate northward [169, 123].

Finally, both statistical analyses of past TC activity [87, 255, 81] as well as model forecasts [1, 155] suggest increasing TC intensity and destructiveness in a warming environment. Studies investigating TC intensity trends from the instrumental record have been highly debated. Some studies concluded that storms have become increasingly more powerful [87, 255, 81]. This can be expected as SSTs increase, more water vapor is generated in the lower troposphere. Subsequently, more energy will be available for TC development, causing storms to be stronger [132, 23]. Other studies, however, have contested that view and found no significant trend in present TC intensity of destructiveness [2, 39]. Increased TC destructiveness can also be attributed to increase coastal populations in recent decades [2]. High resolution model simulations, however, have predicted an increase in overall storm intensity (although simulated TCs tend to be much weaker than observed storms) [155] as well as an increase in the proportion of the most intense storms [20, 19].

From the motivations above, it is clear that there is a feedback loop between climate and TCs. As climate changes, it is likely to impact TC activity and as TC activity changes it will likely impact climate. Unfortunately, in addition to a lack of theory, climate models cannot adequately model some of these climate-TC feedback [222]. Furthermore climate model data output are too coarse to properly model cyclogenesis [88]. Thus, until models are able to output data at proper resolutions (1km or less), we must rely on proxy numerical methods that best describe the climatology of cyclogenesis (rather than its exact physics and thermodynamics).

Climate science has become a popular avenue for computer science contributions given some of the pressing challenges highlighted above, in addition to the recent explosion in climate data availability. Although questions continue to surround the data's quality and reliability [42, 165], we see multiple opportunities for high impact research using data-driven approaches. From a numerical perspective, applying statistical learning methods to climate data has been challenging because the most elegant models impose assumptions on both the processes generating the data (Gaussian, Poisson, *etc.*) or on the data themselves (independence, independent and identically distributed or i.i.d., *etc.*) both of which seldom hold in the highly variable and auto-correlated climate data. Another need for extreme event forecasting is to objectively quantify the uncertainty of forecasts, for without a clear measure of uncertainty, results can be misinterpreted or manipulated. Finally, the data's dimensionality presents challenges in light of the low-sample of extreme events, where the data's dimensionality d is significantly larger than the number of observations n ($d >> n$).

Recent studies investigating Atlantic ocean basin-wide trends of SST and annual

Atlantic TC counts concluded that the high TC frequency cannot be attributed to an increased SST [255, 235]. These studies explained that since the Indian and Pacific basins experienced a similar upward SST trend while the total number of TCs in those basins decreased, higher SST alone could not explain the increased TC activity. Similar studies investigating TC trends usually view TC activity through a basin-wide lens and investigate how changes in climate within these basins affect TC activity. We ask, however, are SST basin averages a good metric to study TC activity?

Figure 4.2 shows the Atlantic basin's mean-adjusted SST and TC counts between 1982-2009. Although the time-series are fairly correlated (0.44), the region's sheer size makes it difficult to formulate a relationship, especially since different regions in the basin might be subjected to different climate phenomena thus potentially impacting TCs differently. Averaging across an entire basin could, therefore, suppress useful information.

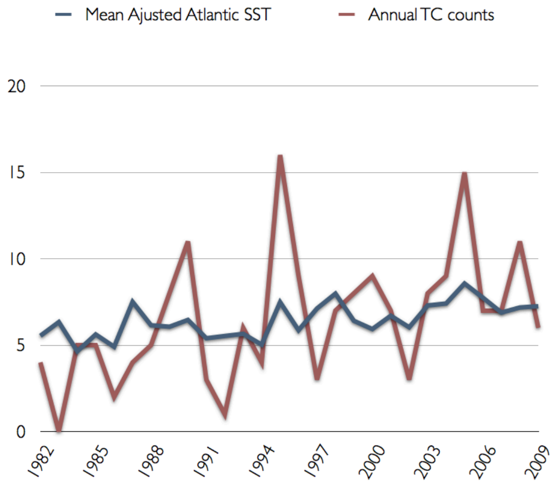


Figure 4.2: **Suppressing information.** Mean-adjusted Atlantic basin SST and annual TC counts for 1982-2009 (0.4367 correlation). The same (thin lines) data from Figure 4.1 superimposed for comparison. We subtracted the SST mean and divided by the standard deviation to scale the SST data to fit the storm counts time-series. Averaging over the entire basin might suppress useful information about the relationship between TCs and SST.

The present study attempts to identify the relationship between Atlantic SST and Atlantic TC activity; and if such a relationship can be captured, how does it evolve over time and how does it impact future TC frequency? The current work provides three new data-driven understandings with regards to Atlantic SST's role in TC frequency:

1. Smaller regions in the Atlantic provide better alternatives to studying SST and TC trends than basin-wide analyses.

2. The warming of Atlantic SST near $20° - 30°$ N *prior* to TC seasons as well as warming of the path along $12° - 15°$ N and $18° - 60°$ W during TC season has an impact on increased tropical Atlantic TCs.

3. Unlike other basins, Atlantic TC frequency could be related to the increase in Atlantic SST.

We begin by reviewing the studies investigating the SST-TC relationship. We will also provide the reader with a brief discussion on the quality of Atlantic TC data and their implications to attributing changes in TC activity to changes in SST. We then introduce a spatio-temporal method that allows to identify more targeted oceanic regions and their relationship to Atlantic TCs. We end the chapter with a discussion of our findings and their implications.

4.2 Review of TC-SST studies

We start by reviewing the relevant literature investigating the SST-TC frequency relationship. Overall, we show that the majority of works are of either anecdotal nature or heavily rely on averaging over large spatio-temporal ranges. Such approaches have limited the insights these studies brought fourth, as demonstrated by our meager understanding of various factors influencing cyclogenesis [86, 195] and the great disparity amongst simulated TC frequency projections [155].

Palmn [191] established that tropical cyclones form only over ocean water with temperatures greater than 26°C. Fisher [101] tracked 11 hurricanes between 1953 and 1955 and concluded that storms tend to form and intensify over warm waters and dissipate over significantly colder water. Gray [119] showed that genesis only occurs in environments characterized by small vertical shear of the horizontal wind and also favors regions of relatively large cyclonic low-level vorticity. He later established a set of necessary conditions (though by no means sufficient) for genesis [114]. In addition to the two aforementioned factors, Gray argued that larger values of the Coriolis parameter, the heat content of the upper ocean (reflecting the depth of the ocean mixed layer), and the relative humidity of the middle troposphere all favor genesis. Carlson [33] attempted to relate the SST along the path of African disturbances to the main development region (MDR) by comparing the August mean SSTs from 1965-1969. He observed that active years experienced the highest SST and low cloud cover over the eastern Atlantic, while the most inactive years analyzed had the coolest SSTs and had greater cloud cover. In similar anecdotal study, Shapiro [213], correlated the first EOF mode of Aug-Oct hurricane frequency with May-July sea level pressure (SLP) to avoid the question of causality (did the hurricanes influence large-scale patterns or vice versa). Low MJJ SLP had -0.5 correlation with ASO hurricanes counts. Low SLP coincided with active hurricane seasons. The author similarly analyzed the correlation of variance normalized MJJ SST ($10 - 45°$ N - $110 - 0°$ W) from 1899-1967 with ASO hurricanes. The highest correlation was 0.55 in the MDR near the west African coast (as in [33]). The author argues that that hurricane activity is sensitive to the SST of that region given that it normally experience low SSTs (on average below the 26.5 "threshold") associated with upwelling off the African coast. It is concluded that higher SST off the region coincides with high hurricane activity but due to inter-correlations between SST and SLP, SST provides little predictive skill. Goldenberg *et al.* [112] used an EOF analysis on the non-ENSO SST mode (removed ENSO effect on SST first), which represents the interannual to multidecadal variability (refereed to the Atlantic multidecadal mode or AMM). The authors find that 5-year smoothed annual means of the AMM in North Atlantic (40 to 70N) has a 0.72 correlation with 5-year smoothed major hurricanes in MDR (10 to 14N - 20 to 70W) and 0.81 with 5-year smoothed Net Tropical Cyclone activity (NTC). The author identified cycles of relative high and low Atlantic TC activity: active periods from the late 1920s through the 1960s and again beginning 1995, and inactive periods from the 1900s through the mid-1920s and during the 1970s through 1994. Heightened activity has during these active cycles coincided with a simultaneous increase in SST in tropics and North Atlantic and decrease in vertical wind shear in tropics. Webster *et al.* [255] analyzed basin-wide mean SST and seasonal TC counts in all the major basins between 1970-2005 and concluded that the upward trend in Atlantic TC seasonal counts cannot be attributed to the increased SST. This was because not all basins that had an increase in SST, had a corresponding increase in TC counts. Later, Hoyos *et al.* [136] using mutual information theory to conclude that the trend of increasing numbers of category 4 and 5 hurricanes for the period 1970-2004 is directly linked to the trend in sea-surface temperature; other aspects of the tropical environment, although they influenced shorter-term variations in hurricane intensity, did not contribute substantially to the observed global trend. By visually analyzing hurricane and SST trends, Holland and Webster [131] find that, since 1900, North Atlantic hurricane activity went through three distinct phases separated by sharp transitions around 1930 and 1995; each regime experienced 50% more hurricanes than the previous one. Moreover, the authors concluded that each phase shift coincides with increased SSTs in the eastern Atlantic ($5-25^\circ$ N, $55-20^\circ$ W). The ratio of minor versus major hurricanes in any given year seemed to be connected to the SST of the Gulf of Mexico ($5-20^\circ$ N, $120-90^\circ$ W) and the proportion of hurricanes that form equator-ward of 25°N. Sanders and Lea [208] built a multiple linear regression model using the MDR Aug-Sep SST and the 925hPa u-wind field averaged over 7.5-17.5 N, 30100 W. They then isolated SST's impact on the model's output to conclude that SST was responsible for the increase TC frequency at the 99th percentile. The authors concluded that the sensitivity of tropical Atlantic hurricane activity to AugustSeptember sea surface temperature over $1965--2005$ was such that a 0.5 C increase in SST was associated with a 40% increase in hurricane frequency and activity. The results also indicate that local sea surface warming was responsible for 40% of the increase in hurricane activity relative to the $1950-2000$ average between 1996 and 2005.

As it can be seen from this brief overview, limited insight has be gained in the past 50 years as to SST's impact on TC frequency. On the one hand, the earliest studies, such as [191, 101, 119, 33] were all anecdotal and it would be unreasonable to generalize their findings to all storms. On the other hand, the remaining studies all employed a variation of spatio-temporal smoothing. Shapiro [213], despite computing global pointwise correlations, never discusses the issue of field significance [170] or false discovery rates [21]. In addition to not discussing the significance of correlations, it is impotent to note that Goldenberg *at al.*'s [112] data had only two full oscillations and was too short to test the likelihood of a cycle [4].Similarly, when discussing hurricane trends Webster *et al.* [255] never take into account that the data is not independent. This might lead to over estimating trend significance as traditional tests assume data independence. Furthermore, their conclusion that SST cannot explain current hurricane activity was premature given that different basins could have different driving conditions for cyclogensis. Holland and Webster [131] never conducted any statistical significance testing on their trends, and as pointed out by [4] the three hurricane "phases" are only significant at the $50 - 80\%$ level and that the signal is dominated by the strong SST phases. Furthermore, when the hurricane data is properly visualized the three distinct phases become less obvious. Another important observation is that numerous correlations are reported between smoothed observations while no physical explanation is provided [112, 255]. Finally, when analyzing seasonal trends all studies looked at contemporaneous correlations where SSTs and TCs from the entire season were analyzed. Such a long temporal averaging make it difficult to deconvolution which months contributed to TC frequency.

4.3 Atlantic Data & Attribution

Attributing a change in TC frequency to external forcing is challenging due to questions surrounding the atmospheric data, the inter-annual and decadal variability of storm frequency, and TC historical records.

Although the Atlantic basin is considered near-fully covered thanks to a network of ships, buoys, satellites and aircrafts. Records are not considered reliable prior to the 1940s when aircraft reconnaissance was introduced. Important uncertainties in atmospheric data believed to affect TC formation make it difficult to quantify the impact of these data on past and future TC activity. For instance, even during reliable periods, coverage is still considered incomplete due to sparse airplane routes and lack of records during World War II [88]. Also, various data reconstruction methods produce different centennial scale trends and can produce substantially different regional TC counts when used to force atmospheric models [155]. Furthermore, it is expected that TC activity could depend on how several highly uncertain (especially at the regional level) thermodynamical and dynamical conditions develop [242, 243], thus making future assessments difficult.

For the inter-annual variability of storms, there seems to be two schools of thought when it comes to TC variability [88]. The First, links Atlantic SST to the SST of the Northern Hemisphere and explains the upward trend in both temperatures as a result of radiative forcings, some of which anthropogenic [231, 171, 131]. This view attributes the warming of the past 30 years to greenhouse gases and the cooling in the $1950-1960$ period to increased concentrations of anthropogenic sulphate aerosols. The second, attributes the decadal variability to natural oscillations of the ocean-atmosphere system [112].

Atlantic TC data is considered the most reliable TC data and dates back to the 1800s thanks to dense shipping routes. However, accuracy is contested until satellite era observations. Currently, it is estimated that nearly 70% of TC reconnaissance is obtained through satellite imagery [201]. Prior to the satellite era, storm reconnaissance was through coastal or ship observations. Given a large number of storms never reach land and that ships generally try to avoid storms, a large number of storms might have gone undetected. Another source of undercounts is low coastal population density during the earlier record [165]. Furthermore, it has been suggested that a storm per year undercount as late as the $2003-2006$ period is possible due to changes in data processing methods [165].

To address these potential undercounts, some studies have resorted to investigate post-satellite era trends only [255, 147]. However trends from such studies can be contested as being part natural oscillations and if longer data were used no significant trends could be observed [39, 162].

Other studies sought to correct the undercounts, however, no consensus on how or which periods to correct has emerged from such attempts. One approach assumed that the ratio of landfalling TCs was relatively constant over time and that a decrease in landfall ratio implies an increasing number of unobserved TCs [220, 165]. In addition to assuming a constant landfalling ratio, Solow and Moore [220] assumed that the true basin-wide TC count is generated from a Poisson distribution (see Appendix **??**) and then tried to reconstruct the earlier TC record using Maximum Likelihood Estimation (MLE), although they focused on trend detection rather than exact count predictions. Landsea [165] estimated that nearly 3.2 storms per year were missed for 1900-1965 and 1 TC per year for 1966-2002. Conclusions from such studies are questionable if the ratio of landfalling TCs changed over time. In fact, Holland [131] presented a recent decrease in landfall ratio due to the warming of the eastern Atlantic. Such results could mean that estimates by Landsea [165] have a high bias.

Another approach was to match satellite-era TC tracks with earlier ship tracks and landfalling locations [42, 245]. Chang and Guo [42], investigated whether there were significant undercounts in open-ocean and near-shore (within 300km from land or islands) TCs. To do so, they compared satellite era TC tracks with pre-satellite ship observations to see whether there were a significant number of tracks that did not fall within any ship routes and therefore would have gone undetected during the pre-satellite era. They find that the number of TCs that did not make landfall and went undetected were 10 TCs per decade or less. These results do not comment on the potential change of TC track properties over time, furthermore the authors relaxed the requirement for observed storms from two to a single ship observation. Similarly, Vecchi and Knuston [245] used pre-satellite era ship-track data and satellite-era TC locations to identify adjustments in the number of annual TCs, tropical storm days, average TC duration, and TC density. They conclude that nearly 3.4 storms per year were missed starting in 1880 and about 0.25 storms per year in 1960 with perfect coverage thereafter. This study makes several assumptions, most notably that present-day TC activity is representative of past activity.

Finally, some studies relied on statistical models to infer the number of missed TCs [172, 219, 186]. Mann *et al.* [172] used a Poisson regression model to estimate seasonal TC counts based on the SST in the Atlantic main development region (MDR; see Appendix A), the El-Niño Southern Oscillation index (ENSO), and the North Atlantic Oscillation index (NAO) of the following winter as predictors. To test their model, the data was divided into two intervals ($1870 - 1938$ and $1939 - 2006$) and one interval was used to train and the other to test the data. Subsequently, the authors compared the performance of their model using no adjustment, minor, and major TC undercount adjustments. The results suggest a modest undercount of one storm per year. Results from this work can be questionable, however, given the limited testing of the regression model and the rigid selection process for the training and testing periods. Solow and Beet [219] estimated TC counts using a Poisson regression model with Atlantic MDR SST and ENSO of the following winter as predictors. The results were presented in terms of observation probabilities with 71% in 1870 to 100% by 1964. The authors conclude that these estimates are slightly larger than those of [172]. Nyberg *et al.* [186] reconstructed the past 270 year history of major (category 3-5) hurricane counts by building a neural network model trained on coral luminance intensity, the number of planktonic *Globigerina bulloides* from sediments (paleoclimatic proxy records), SST and vertical wind shear. They estimated 2-3 missed major hurricanes per year during the $1870 - 1943$ period. Mann *et al.* [172] suggested that those estimates translated to nearly 6-20 missed TC per year for the pre-1944 period.

In summary, current TC activity cannot be attributed to global climate change. This is due to a short reliable observational record of atmospheric data, a lack of understanding of how TCs might behave in future climate, and potential TC undercounts that remain unclear as to how and which periods to correct. In the following sections, although I will review work pertaining TC activity in various basins, the main focus will be on the Atlantic given that it has the most reliable historical record.

In summary, a review of previous research pertaining future TC frequency highlights the absence of analytical tools to gain insight into SST's relationship to TC formation and frequency. The present study attempts to quantify such relationship by identifying smaller and more meaningful regions in the Atlantic ocean as opposed to the entire basin, and show that by focusing our attention on critical regions we are able to define a stronger relationship between SST and TC frequency.

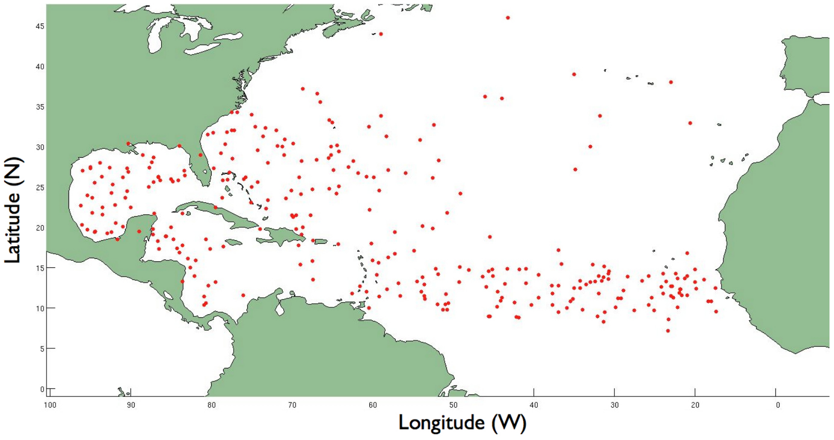


Figure 4.3: **Atlantic Tropical Cyclones.** Location of cyclogenesis for all Atlantic TCs for June-October 1982-2009.

4.4 Terms and Data

4.4.1 Terms used

Tropical cyclones (TCs) are defined as "warm-core non-frontal synoptic-scale cyclones, originating over tropical or subtropical waters, with organized deep convection and a closed surface wind circulation about a well-defined center" [155]. Throughout this paper, tropical cyclones will specifically refer to storms with maximum wind speeds of at least 39 mph and hurricanes will refer to those with maximum wind speeds of at least 74 mph. The strongest hurricanes of type 4-5 will refer to storms with maximum wind speeds of at least 100 mph.

A tropical cyclone basin is a portion of the ocean where storms form. The Atlantic basin refers to the regions in the Atlantic from 90^o to 20^o W and 5^o to 35^o N. The tropical Atlantic basin refers to the region 10^o to 20^o N, excluding the Caribbean west of 80^o W (as defined in [235]). The Atlantic tropical storm season refers to the months known for high Atlantic TC activity. In this case, we consider the months of June-October (as defined in [255]) and for brevity we will refer to these counts as "annual".

4.4.2 Data

We used NOAA's optimum interpolation (OI) SST analysis dataset [203].The data consist of monthly means from 1981-2010[1] on a 1° grid.

For storm counts, we used the Unisys Atlantic Tropical Storm Tracking dataset[2] containing all Atlantic storms from 1870-2009 (Figure 4.3). We parsed all storm counts from the Unisys website and the data is available in .mat format at www.cs.umn.edu/∼jfagh/tc.mat. As far as we know, this is the best Atlantic storm count data publicly available and accessible to non-climate scientists.

[1] The dataset contained partial data for the year 1981, which was not included in the analysis

[2] http://weather.unisys.com/hurricane/atlantic/

4.5 Identifying smaller SST regions in Atlantic that outperform basin-wide averaging

As shown in Figure 4.2 the Atlantic basin average SST has a moderate correlation with TC originating in the entire Atlantic basin (0.44). Although previous work [69, 198] suggested regions of interest within the Atlantic near the African West coast, a systematic search for regions that better explain TC frequency has not been undertaken. Algorithm 1 describes the linear optimization problem we solved to identify SST regions that shed light on the SST-TC frequency relationship.

The objective of Algorithm 1 is to identify an SST and cyclogenesis region in the Atlantic where the SST region's mean annual ($\overline{SST}$) has the highest linear correlation (ρ) with TCs originating from the identified cyclogenesis region.

$$
\begin{aligned}
&\textit{maximize} && \rho(\overline{SST}_{lat,lon,w,h}, storms_{lat',lon',w',h'}) \\
&\textit{subject to} && \textit{minimize}\ \Delta(SST_{lat,lon}, storms_{lat,lon}) \\
&&& 20^\circ \geq SST_{w,h} \geq 10^\circ \\
&&& 45^\circ \geq SST_{lat} \geq 10^\circ \\
&&& 80^\circ \geq SST_{lon} \geq 18^\circ \\
&&& 60^\circ \geq storms_{w',h'} \geq 20^\circ \\
&&& 45^\circ \geq storms_{lat'} \geq 10^\circ \\
&&& 80^\circ \geq storms_{lon'} \geq 18^\circ
\end{aligned}
$$

Algorithm 1: Find the best correlated SST and storm region.

When we run Algorithm 1 on the TC season (June-October) SST mean, we find that the SST region off the African coast centered around 15°N and 20°W and the region in the tropical Atlantic 10-20° N and 18-60° W are the two highest correlated SST-cyclogenesis regions of the entire Atlantic basin (Figure 4.4). The highlighted cyclogenesis region accounted for nearly 40% of all Atlantic tropical storms and nearly 50% of all Atlantic hurricanes during the period we analyzed (see Figure 4.7). To compare our region to previously used methods, we computed the correlation of mean Atlantic SST with all TCs, hurricanes, and type 4-5 hurricanes that originated in the entire Atlantic basin. Additionally, we computed the correlation of mean tropical Atlantic SST and tropical Atlantic TCs, hurricanes, and type 4-5 hurricanes. Figure 4.5 summarizes the correlation results. Our optimal SST-cyclogenesis region has improved correlations of 0.66, 0.7, and 0.53 respectively versus Atlantic basin-wide averaging (0.44 0.46, and

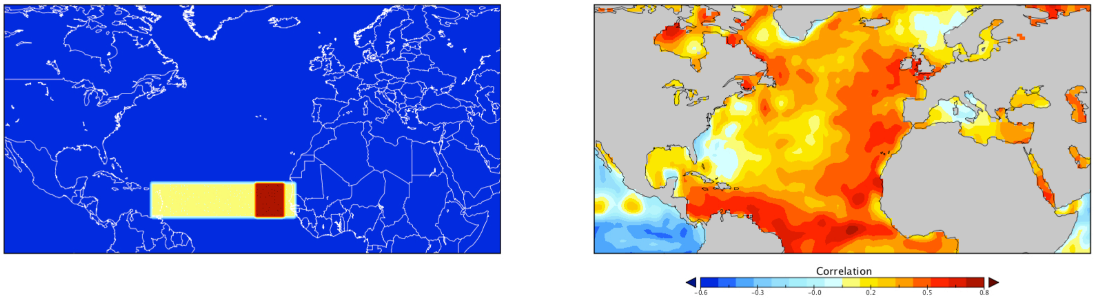


Figure 4.4: **Optimal SST-Cyclogenesis Regions.** *Left:* The optimal SST (red) and cyclogenesis (yellow) regions identified by Algorithm 1. The June-October mean SST from the red box has a 0.66 correlation with the June-October TC counts from the yellow rectangle, compared to 0.44 for basin-wide Atlantic SST and TCs. *Right:* the linear correlation between mean annual SST and TC counts originating from the cyclogenesis (yellow) region identified on the left panel. Notice other highly correlated regions with TC counts. The output of the optimization algorithm depends on the constrains imposed on the problem (size, distance, *etc.*)

0.43 respectively). The sharp decrease in correlation with the most intense hurricanes in Figure 4.5 is to be expected as the SST at cyclogenesis is a poor predictor for future storm intensity [84].

We must note that the output of Algorithm 1 depends on the constraints imposed on the optimization problem. As it can been seen from Figure 4.4 (right panel), regions other than the red box in Figure 4.4 (left panel) correlate highly with TCs originating from the yellow region yet they were not selected as an optimal pair. The reason why some regions are selected over others is the constraints imposed on the optimization function. In our case, we imposed that: *(i)* distance between the two regions be minimized; *(ii)* Both boxes must be at least 10° by 10°; and *(iii)* both boxes must be north of 5° N. Removing any of these constraints would result in different optimal SST-cyclogenesis regions. While it is plausible that far away SST and cyclogenesis regions might impact one another (*i.e.* teleconnections), the focus of this paper is on local SST impact on TC frequency therefore we constrained the problem to nearby SST and cyclogenesis regions. We only examined regions that were at least 10° (in either dimension) to avoid any spurious correlations between extremely small regions. Finally, we examined locations north of 5° N given that cyclogenesis is only possible at a certain distance away from the equator (see also Figure 4.3).

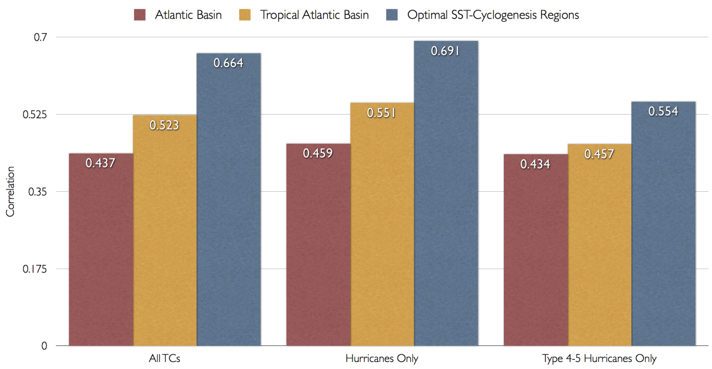


Figure 4.5: **Improved correlation.** Linear correlation scores of regional SSTs and TC counts for Atlantic Basin (5^o to 30^o N and 20^o to 90^o W), Tropical Atlantic Basin (10^o to 20^o N, excluding the Caribbean west of 80^o W), and the optimal SST-cyclogenesis region identified by Algorithm 1 (see Figure 4.4). The optimal SST-cyclogenesis region correlates better with local TCs than both basin-wide averages.

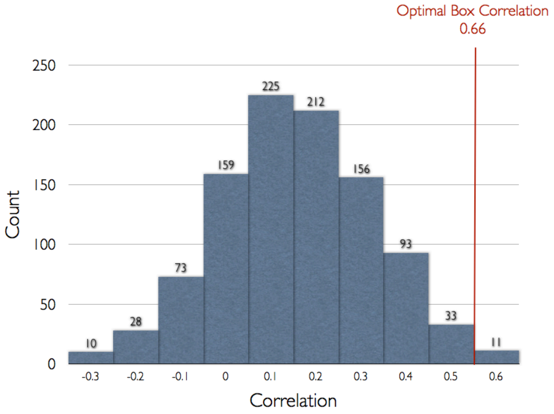


Figure 4.6: **Non-Random Correlation.** Linear correlation scores of optimal SST-cyclogenesis regions with random ordering of annual storm counts. The 0.66 correlation of the regions identified in Figure 4.4 is greater than 99.8% of correlations in the randomization test. We can conclude that correlation between with two optimal regions in Figure 4.4 is not random.

In order to ensure the correlations between the regions identified in Algorithm 1 aren't spurious, we ran a randomization test ($N = 1000$) where we shuffled the years in which the individual storms occurred and ran Algorithm 1 for that random permutation. Figure 4.6 shows the correlation scores of the 1000 optimal SST-cyclogenesis regions identified based on the randomized storm counts. As the correlation between the SST-cyclogenesis regions identified in Figure 4.4 is 0.66, which is greater than 99.8% of the correlations in the randomization test, we conclude that the optimal region's correlation is not random (only two tests resulted in correlations greater or equal to 0.66).

Given that the majority of TCs originating in the tropical Atlantic are of African easterly origin [163, 112], the optimal SST-cyclogenesis regions are physically consistent as well. African easterly waves (AEWs) consist of an elongated area of relatively low air pressure, and travel from east to west across the tropics causing increases in clouds and thunderstorms. They form as a result of the pressure gradient between the higher pressure in the subtropics and the low-pressure equatorial region. AEWs have been identified as integral part of West African and North Atlantic climate [135]. AEWs traveling westward from the West African coast and along the tropical Atlantic (also known as the main development region) tend to intensify and transform into TCs and intense hurricanes [112]. Therefore, our results suggest that the warming of the Atlantic around $10° - 20°$ N and $18° - 60°$ W might provide additional energy for AEWs to develop into TCs.

When we further analyze the cyclogenesis region identified by Algorithm 1 (which lies within the tropical Atlantic), we find that the TC activity of the two tropical regions in the Atlantic east and west of 60° W accounts for nearly 70% of all Atlantic TCs 4.7. Interestingly, the TC frequency east and west of 60° W are anti-correlated (-0.56 Figure 4.8, left panel) suggesting that TCs from each region have different climatologies, and therefore averaging over the entire Atlantic basin would suppress singular TC characteristics. When we compare the two TC frequencies to El-Niño Southern Oscillation (ENSO) index [258], a coupled ocean-atmosphere phenomenon linked to global climate variability, we find that it is anti-correlated with TCs originating east of 60° W (-0.27) and correlated with TCs originating west of 60° W (0.26). The El-Niño phase has already been linked to suppressed Atlantic TC activity [115, 229], it is unclear however if ENSO has a direct or auxiliary effect on Atlantic TCs. When we examine Figure

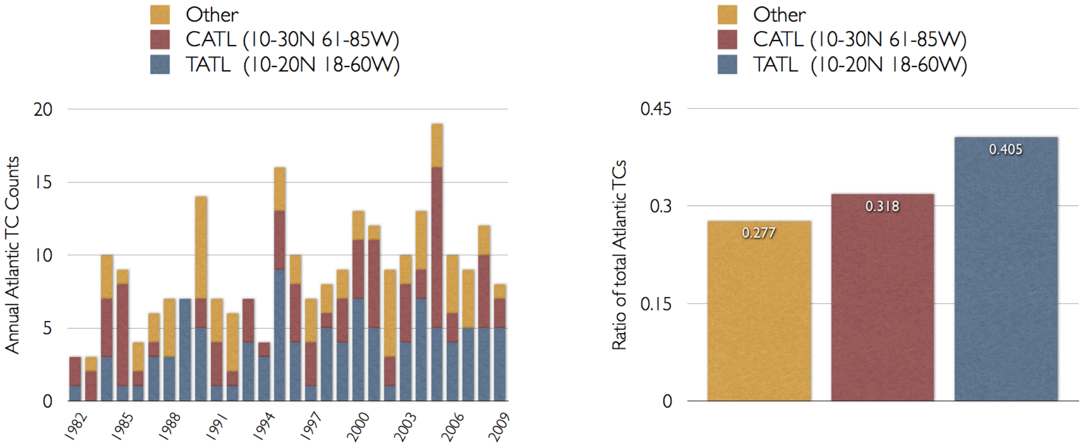


Figure 4.7: **Where Are Atlantic TCs Originating From?** *Left:* Annual Atlantic TC counts grouped based on cyclogenesis region. We divided the Atlantic basin into the tropical Atlantic (TATL), Caribbean (CATL), and the rest of the basin (Other). The figure indicates that when either the tropical Atlantic or Caribbean has an active season, the other experiences relatively low activity. *Right:* The mean percentage of total Atlantic storms that originated in the three regions monitored. Together the tropical Atlantic and Caribbean account for nearly 70% of all Atlantic TCs.

4.8, we notice basin-wide low activity seasons coincide with El-Niño years (2002, for example). However, the percentage of TCs originating east of 60° W rises during some El-Niño years (1995, 1997, and 2009). We conclude that although El-Niño might have a basin-wide suppressive effect on Atlantic TCs, the percentage of TCs originating from a given region might rise or fall depending on the phase of ENSO.

In summary, this section demonstrates how focusing our attention on smaller regions provides better insight into the SST-TC relationship. Algorithm 1 is able to identify an SST region that lies along the path of AEWs, a primary source of TCs and intense hurricanes, and correlates highly with tropical Atlantic TCs. Furthermore, dividing the TCs into groups based on their cyclogenesis region showed that TC frequency in the tropical Atlantic and the Caribbean are anti-correlated and that El-Niño has a more pronounced effect on Caribbean TCs rather than all Atlantic TCs.

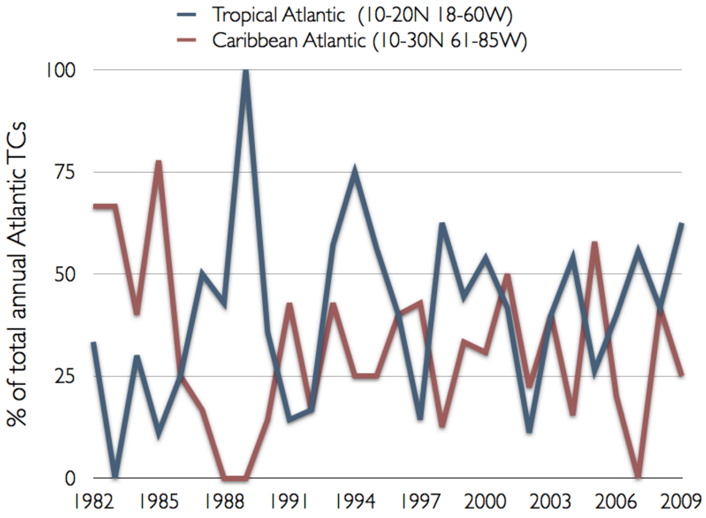


Figure 4.8: **Tropical Atlantic and Caribbean TC Frequency.** Percentage of Atlantic TCs originating from the region identified by Algorithm 1 (tropical Atlantic $10° - 20°$ N and $18° - 60°$ W) and the Caribbean ($10° - 30°$N and $61° - 85°$W). The two activities are anti-correlated at -0.56 suggesting that TCs in each region have different climatologies, therefore averaging across the entire Atlantic might suppress useful information.

4.6 TC precursors

While factors influencing TC activity other than SST have been identified [119, 115], it is useful to understand how SST prior to the hurricane season might affect TC activity. To do so, we applied Algorithm 1 to the May-June mean SST and annual TC counts. Figure 4.9 (left panel) shows that the May-June SST in the red box has the highest correlation with June-October TCs originating from the yellow rectangle (0.55). This correlation suggests that warmer SSTs near $20° - 30°$ along the West African coast precede active TC seasons (Figure 4.9 right panel). However, given that TCs operate at timescales of days or hours, it is unlikely that May-June SST is a direct cause of seasonal TC frequency. Instead, it seems that the warming of the North Atlantic off the African West coast might provide favorable conditions for other factors that induce tropical cyclogenesis.

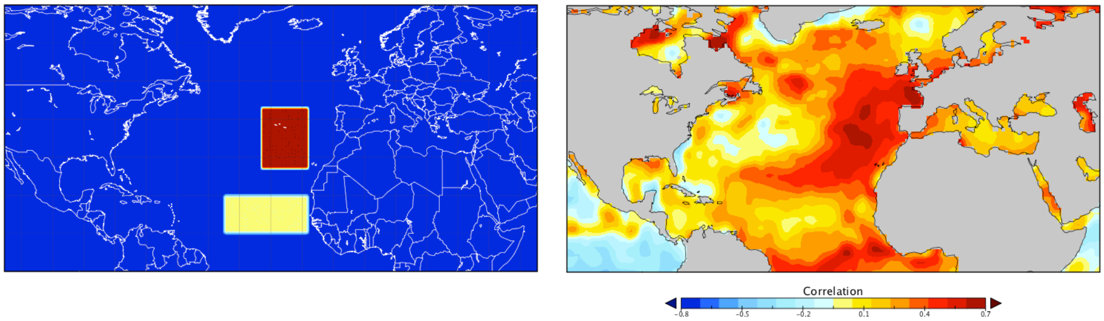


Figure 4.9: **TC precursors.** *Left:* The optimally correlated May-June SST region (red box) and annual (June-October) TC counts (yellow rectangle) identified by Algorithm 1. *Right:* Linear correlation heat-map for May-June mean SST with June-October TC counts originating in the yellow rectangle (left panel). The warming of the region north of $20° - 30°$ N precedes active TC seasons.

As stated in the previous section, most tropical Atlantic TCs are of AEW origin. Although the number of AEWs is relatively constant from year to year [104], the percentage of developing AEWs is highly variable [112, 232]. Recently, Sall *et al.* [206] have shown that, at the end of their continental path, most AEWs weaken and vanish over the ocean, while only a few of them strengthen and become tropical storms. This seems to be the case even when the departing perturbation appears well developed and the large scale environment is favorable for tropical cyclogenesis. A recent numerical

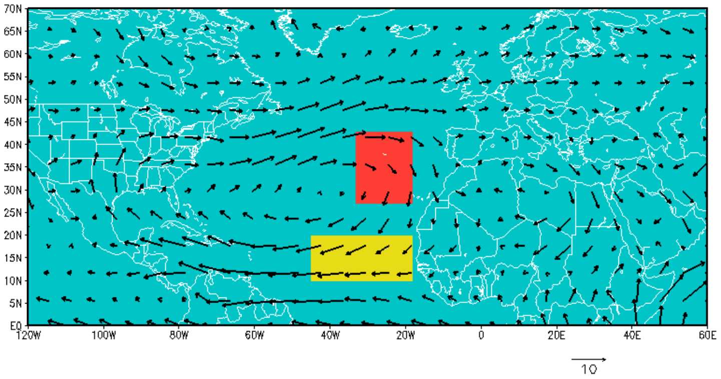


Figure 4.10: **Warm SSTs Traveling Southward.** Mean May-June low-level (859 hPa) wind fields for the 1982-2009 period. The movement of the wind during the period of investigation confirms that it is plausible that warmer SSTs migrated southward and provided additional moisture to the cold dry air traveling from higher latitudes, effectively reducing the chances of suppressing AEWs exiting the West African coast. The arrow length denotes 10 meters per second.

modeling study showed that the northerly isentropic descending airmass below the Saharan Air Layer traveling from the drier middle troposphere at higher latitudes to the lower troposphere at lower latitudes provide a substantial source of dry air off the West African coast and effectively disrupts cyclogenesis [69, 68]. More specifically, AEWs exiting from the African continental landmass tend to ingest dry air descending from middle latitudes (Europe and North Africa) and dissipate without experiencing cyclogenesis. These findings are consistent with recent reanalysis studies that tracked and analyzed AEWs in the ERA40 reanalysis data and found that non-developing AEWs experience dry mid-to-upper-level air [135]. Figure 4.10 shows the mean May-June wind circulation for 1982-2009. We can see that wind motion is consistent with our hypothesis as the wind transports the warm air (red box) from higher latitude towards $10° - 20°$N (the yellow rectangle).

Subsequently, our analysis shows that the warming of the region off the West African coast (red box in Figure 4.9) in May-June, would provide additional moisture to counter the dry air descending from higher latitudes and therefore increase the chance that AEWs advance deep into the Atlantic and develop into TCs. Moreover, it seems that

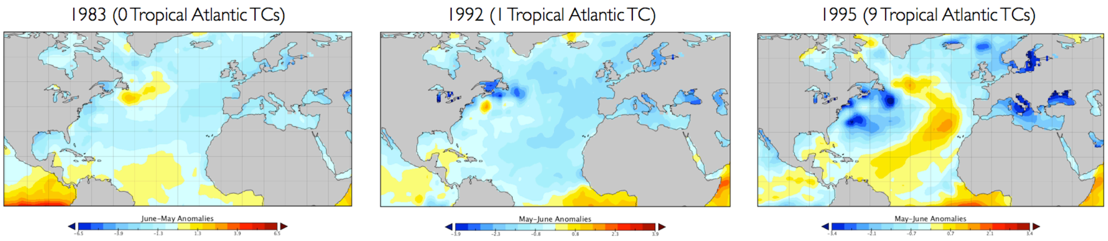


Figure 4.11: **May-June SST Anomalies.** *Left and Center*: Negative May-June SST anomalies north of AEW exit region preceded low tropical Atlantic TC seasons (1983 and 1992 with 0 and 1 TCs respectively). Notice in the leftmost figure the high anomalies in the Pacific (lower left corner) supporting the results of negative correlation between El-Niño and tropical Atlantic TCs. *Right:* Positive May-June SST anomalies preceded one the most active Tropical Atlantic TC season on the record (1995).

substantial warming (or cooling) of that region could have a pronounced effect on Tropical Atlantic TC frequency. When we examine two abnormal TC seasons where the optimal cyclogenesis region (yellow rectangle in Figure 4.4) generated a low number of storms, we see how May-June SSTs might have an effect on June-October TC frequency. In these two instances (left and center panel in Figure 4.11), the region identified previously showed abnormally low SSTs in the months prior to TC season and subsequently had its least active TC season in the data we analyzed. While, May-June anomalies could not alone explain low and high activity TC seasons in the tropical Atlantic, our analysis suggests that when the region is cooler than average it cannot counter the dry air descending from higher latitudes and subsequently AEWs exiting the West African coast around $10° - 20°$N are suppressed. It important to note that although Figure 4.11 highlights instances where high/low May-June anomalies precede high/low TC frequency. May-June anomalies alone cannot explain annual TC frequency as there are years that have significant negative May-June anomalies yet normal TC seasons. However, monitoring May-June anomalies in addition to other TC precursors can provide a better understanding of the SST-TC frequency relationship.

4.7 The Atlantic SST-TC relationship

The previous two sections presented additional insights that help us understand the Atlantic SST-TC relationship better than with basin-wide averaging alone. Figure 4.1

shows upward trends in Atlantic SST and TCs similar to those in [255, 235]. Although [255, 235] did not link the increased Atlantic TC frequency to SST, we propose that the lack of increase in storm numbers in other basins experiencing increased SST is not due to the absence of a significant relationship between SST and TC frequency. Rather, it is due to the absence of another factor influencing cyclogenesis namely, AEWs, which may have a more pronounced effect on cyclogenesis than their Pacific counterpart. On the one hand, Pacific easterly waves (PEWs) are driven primarily by convective heating, which depends on SST. On the other hand, the barotropic to baroclinic conversion, which is the energy transport from the mean flow toward the rotational flow component, dominates AEWs. This means that AEWs are strongly associated with rotation and therefore cyclogenesis, unlike PEWs, which are related to convection (i.e. the vertical flow) [212]. In summary, we maintain that prior to- and during the TC's formative phases, the warming of $20° - 30°$N along the West African coast as well as along the path of AEWs has a significant impact on whether disturbances develop into TCs. Thus, warmer SSTs off the West African coast prior to TC season and along the AEW path during TC season would provide a more favorable environment for cyclogenesis and could be a reason for an increase in annual Tropical Atlantic TC frequency.

4.8 Conclusion

4.8.1 Summary

In this chapter, we reviewed the major works that investigated the SST-TC relationship. We demonstrated the ability to apply data-driven techniques to observational climate data to supplement physics-based models. We showed that data-driven methods can help identify smaller regions in the Atlantic ocean that provide a better glimpse into the relationship between Atlantic SST and Atlantic TC frequency compared to basin-wide averages. Our results showed that when investigating the relationship between SST and TC frequency, it is better to look at smaller regions, or centers of action, instead of basin-wide trends. Coincidentally, the regions identified in this study are consistent with existing theory about AEWs and TC development. Specifically, the SST of the region north of 20^o N off the West African coast prior to TC season as well as that along the AEW path (10°-15° N and 20°-60° W) during TC season have significant impact on

Tropical Atlantic TC frequency compared to other regions in the Atlantic. Furthermore, an increase in both mean Atlantic SST and TCs can be attributed to the warming of the regions we highlighted above in conjunction with AEW activity. In similar fashion, if we monitor TCs based on their cyclogenesis region, we find that the percentage of TCs originating from the tropical Atlantic east of 60°W is anti-correlated with that in the Caribbean. This suggests that different climatologies govern each TC region, and are potentially mutually exclusive. Finally, when we compare these cyclogenesis percentages with the El-Niño index, we find that, contrary to previous assumptions that the Atlantic TC frequency is suppressed when El-Niño is present, some regions in the Atlantic, most notably the Caribbean is actually positively correlated with El-Niño. All of these results support our claim that in order to gain proper insight into the SST-TC relationship basin-wide generalizations are not sufficient.

4.8.2 STDM Contributions

We developed a scalable spatio-temporal relationship mining tool capable to looking for both linear (correlation) and non-linear (MIC) relationships. Despite the large search space, our open-source tool *QuickSig* is able to compare millions of relationships between arbitrary-sized regions in seconds. We also provide various randomization significance tests such as bootstrapping and block randomization. The tool is generalizable to any set of spatio-temporal variables.

4.8.3 Discussion and Future Work

Based on these findings, future work investigating SST's impact on TCs should not indiscriminately average basin-wide temperatures. Instead, critical regions must be identified and treated as such. Our approach suffers from a few limitations. First, it heavily relies on linear correlations and therefore cannot capture non-linear relationships between SST and TCs. Additionally, some Atlantic TCs do not generate in the Tropical Atlantic (North Atlantic storms in the region north of 30° N, for instance) and to fully understand the Atlantic SST-TC relationship we must understand the climatic conditions that govern those TCs as well.

Many questions remain after this analysis as we continue to face several challenges described in Chapter 2.0.2. Specifically, the highly autocorrelated SST field makes significance tests extremely challenging as any one time-series that may correlate highly with TCs by random chance will likely mean that its neighbors will have an artificially high correlation. Second, it is impossible to tell with certainty whether SSTs are a driver or a mere response to other large-scale phenomena impacting TCs. For instance, the north Atlantic region identified when analyzing pre-season SST is also associated with a large-scale phenomenon known as the North Atlantic Oscillation (NAO), that may be driving TC activity. Finally, although there seems to be an upward trend in SST and TCs, if de-trended by a 10-year mean many of the high correlations drop significantly. Raising the question whether the high correlation is due to a real relationship or a co-occurring trend.

Chapter 5

S-ENSO

In the previous chapter, we presented an extensive relationship mining application between SSTs and seasonal TC counts. In this chapter, we focus on a more specific type of relationship mining: predictive modeling. Predicting seasonal TC counts is an active field of research with interests to society, the reinsurance industry, and scientific community. So far, our success at predicting season TC activity several months in advance has been limited. In this chapter we present further evidence that monitoring the spatio-temporal evolution of climate variables can improve the accuracy of predictive models. Specifically, we show how monitoring the spatio-temporal warming patterns in the Pacific significantly outperform models that relay on traditional fixed-region variables. We begin the chapter with a broad overview of the state-of-the-art predictive modeling for hurricanes. We then include a brief discussion of the various evaluation methods used to test such predictive models. We then introduce our new spatio-temporal index Spatial ENSO (S-ENSO) and demonstrate how it out performs traditional (non-space-time) indices. We end the chapter with a discussion on the possible mechanisms that lead to S-ENSO's performance.

5.1 Review of TC Forecast Models

Numerous factors have been identified to influence formation and intensification of TCs. Some of the factors are local such as SST [112], vertical wind shear [**?**], sea-level pressure (SLP) [164], Emanuel's unit-less parameter χ_m [88], lower tropospheric moisture [164], wave energy accumulation [252], and African easterly waves (AEWs) [134]. As well as remote factors, which are dominated by climate indices – diagnostic tools used to describe the state of a climate system using composite signals such as the to pressure differences at two locations in the Pacific Ocean – these include: ENSO [229], the Atlantic multidecadal oscillation (AMO) [148], The Madden-Julian oscillation (MJO) [158, 148],and the Quasi-Biennial Oscillation (QBO) [115]. Other remote factors include African dust aerosol [93], Sahel rain fall [267], and solar radiation [83].

In this section, I review the studies that attempt to predict future TC behavior from centennial to weekly timescales using the above predictors. Broadly speaking, we can group TC predictions into three categories: (i) Centennial projections: these simulations are used to model TC activity under various warming scenarios and generally look at TC activity beyond the 21st century. (ii) Seasonal forecasts: forecasts of seasonal TC activity are issued in December (for the Atlantic) the previous year and are periodically updated throughout the TC season. (iii) Short-term forecasts, these forecasts are issues 7-14 days before TC genesis and generally predict intensity and tracks instead of genesis.

5.2 Centennial TC Projections

In order to understand future climate activity under various warming conditions, large-scale global and regional physics-based models are heavily used given that they generally do not rely on observational data. Globally, the majority of climate models project a decrease in annual TC counts ranging from 6 to 34%, with a notable decrease in Southern Hemisphere TCs [155]. Some studies have attributed the reduction to increased vertical wind shear, which is considered negatively correlated with TC formation [119]. Garner *et al.* [108] found an average wind shear increase of 10% in the Atlantic's main development region (MDR). Vecchi and Soden [242] projected a 10% increase in wind shear per degree of global warming. Additionally, Nolan and Rappin [185] find that TC genesis may become more sensitive to wind shear in a warmer climate, which means significant increases in wind shear might not be necessary to curb TC activity. Vecchi and Soden [242], however, found that the contribution of vertical shear to cyclogenesis is comparable to that of other environmental factors (vorticity, relative humidity, and Emanuel's wind Maximum Potential Intensity). Other studies have attributed reductions in global

TC frequency to increased atmospheric stability [227, 44, 188, 266, 20, 120] and the weakening of the tropical hydrological cycle [227, 20] which is associated with a decrease in the upward mass flux accompanying deep convection [125]. Other work [88, 90] suggests that decreases may be due to an increase in the saturation deficit of the middle troposphere (see equation 2 in [88]).

What's most notable about these global projections is that despite consistent projected decrease in total number of TCs, the IPCC 4th Assessment Report gives these predictions "medium confidence" while it gives the more divergent intensity projections [90] "high confidence" [231]. This might be because there is a general consensus regarding the theoretical upper bound on TC intensity [23], whereas there is no consensus concerning cyclogenesis theory.

At the regional level, and for the Atlantic basin in particular, the majority of high resolution simulations are able to reproduce past/current Atlantic TC activity and overall interannual variability (see Figure 3 in [155]). Recent lower resolution simulation (0.9° Gaussian grid) have not reproduced interannual variability, yet it is unclear if the better fidelity in high-resolution models is due to higher modeling resolutions or other model parameters [166]. Aside from these hindcasting experiments, regional climate models (RCMs) agree on very little else. RCMs have a large variation in projected Atlantic TC frequency from -50% to + 30% [120, 188]. When considering all basins, the projected changes for individual basins range up to $+/-$ 50% or more [155]. Not only there is no clear consensus on future Atlantic TC activity, similar models also provide contradicting results. Climate models running over observed SST project a 30% increase in Atlantic TCs due to high temperatures in the Atlantic MDR [188]. Other studies using observed SST, however, projected decreases in Atlantic TC frequency because the minimal SST threshold for cyclogenesis increased at a higher rate than SST warming [153]. Global coupled climate models – models that also consider interactions between the atmosphere and SST instead of only representing SST's influence on atmosphere – forecast either no significant changes [20] or considerable deceases ($\sim -50\%$) under the most aggressive warming scenarios (four times current CO_2 levels). Finally, some statistical/dynamical models show a slight increase (+2.2%) [90] or decrease (-1%) [154]. These disagreements have been attributed to uncertainties in the large-scale patterns of future tropical climate change, as evident in the lack of agreement between the model projections of patterns of tropical SST changes [268] as well as remaining limitations in the downscaling strategies these models use [155]. Additionally, RCMs tend to suffer from spectral nudging that forces them to follow the climatic conditions that drive the large scale model [155].

In summary, physics-based climate models project a global decrease in the total number TCs yet are highly uncertain in individual basins. These uncertainties stem from three major roadblocks: First, the models' output dependence on how well they simulate the large scale conditions that affect cyclogenesis as well as sensitivity to parameterization [152, 166, 268]. Second, global climate model output data are too coarse (20-120km) to fully model TC properties and it is estimated that a grid resolution of 1km or less might be needed instead [49]. Finally, even if it were possible to perfectly model future climate, limited understanding of how these changes would affect TC activity make future projects extremely challenging [195, 88].

5.3 Seasonal Forecasts

Seasonal basin-wide activity predictions of seasonal TC activity are issued as early as April in the previous year (to forecast activity in August-October the following year). Forecasts are generated by both physics-based and statistical models. Similar to the previous section, dynamical models predict the state of future climate and the response of the TC-like vortices in the models is used to estimate future hurricane activity [246, 248, 268, 244]. An approach analogous to model simulations is the statistical approach, where one infers relationships solely based on observational data [115, 77, 149]. depending on the approach used some of the forecasts are either probabilistic (*i.e.* % chance of a below-, above-, or average season) or deterministic (*i.e.* 10 landfalling hurricanes).

For dynamical models, Vecchi *et al.* [244] used a statistical-dynamical model to forecast the Atlantic seasonal TC activity for the 2010 season. The statistical model was a Poisson regression model conditioned on Atlantic MDR SST and global tropical SST. The dynamical model used to generate SST observations was the NOAA High-Resolution Atmospheric Model (HiRAM-C180). In the training phase, the model was able to reproduce the overall interannual variability of TC activity with a correlation of 0.76 and a RMSE of 1.99 hurricanes for the period 1982-2009. For the 2010 season (test), the model predicted an active season compared to the 1982-2009 climatology with a 50% probability of storms between 6 and 9 and an expected number of 8. Similar dynamical models include [246, 248].

Zhao *et al.* [268] used the NOAA High-Resolution Atmospheric Model (HiRAM-C180) dynamical model to predict (hindcast) Atlantic and East Pacific Activity. The predicted TC activity correlated by 0.69 in the Atlantic and 0.58 in the Eastern Pacific. When the model was forced with observed SST (HadISST) the correlations improved to 0.78 and 0.65 receptively. The reduction in skill has been attributed to the model's simulation of the different between MDR SST and the tropical mean SST. To confirm the importance of the tropical mean SST, a linear regression model was built using the observed tropical mean SST and it performed similarly to the full dynamical model (r= 0.55; r=0.62 respectively).

Some of the earliest Atlantic TC statistical forecast models [115, 116] relied on ENSO, the Quasi-Biennial Oscillation (QBO) and Caribbean basin sea-level pressures, all considered remote factors in Atlantic cyclogenesis. Gray and colleagues continued to issue annual forecasts, however, few were published in peer-reviewed literature. Gray *et al.* [118] used a least absolute deviation (LAD) regression model conditioned on November-September QBO and rainfall in the West Sahel (August-September) and Gulf of Guinea (August-November) regions for a 0.44 correlation with climatology. Klotzbach and Gray [150] tried to focus on TC counts for the month of September (the season's most active month) by selecting wind, sea level pressure and geopotential height from various regions on the globe.

The predictors (see Appendix B) were selected using a mixture of top-10 bottom-10 composite analysis where various meteorological parameters such as sea surface temperature, sea level pressure, and zonal wind for the month of active and inactive Septembers, and subsequently select the top-10 and bottom-10 anomalies. Other variables were selected using the NOAA correlation tool[1]. To test their approach the authors compared the hindcast to predicted net tropical cyclone (NTC) activity – an aggregate measure of the following six parameters normalized by their climatological averages: named storms (NS), named storm days (NSD), hurricanes (H), hurricane days (HD), intense hurricanes

[1] `http://www.cdc.noaa.gov/Correlation`

(IH), and intense hurricane days (IHD). In a similar study, Klotzbach and Gray [151] did not use the rainfall predictor used previously in [118] and instead relied on variables similar to [150]. The predictors were found to be all associated with ENSO, the Arctic Oscillation (AO), NAO, the PacificNorth American pattern (PNA), and the QBO. This method showed similar skill to that in [118] (0.66 correlation to actual NTC) despite having longer and more homogenous data.

Balke and Gray [25] use a set of twelve predictors to forecast August NTC, which has been shown to have a significant relationship to August U.S. TC landfalls. The linear regression model was able to capture 55% of the variance of the hindcast data. It's predictive power however was limited with a RMSE of 5.12. The most prominent predictive signal was the July 200-mb wind off the west coast of South America. When it was anomalously strong from the northeast during July, Atlantic TC activity in August was enhanced. Other July conditions associated with active Augusts include a weak subtropical high in the North Atlantic, an enhanced subtropical high in the northwest Pacific, and low pressure in the Bering Sea region.

Nyberg *et al.* [186] used reconstructed proxy records from corals and marine sediment core to predict past hurricane activity. A back-propagation neural network was then trained using the proxy records to predict August-October major Atlantic hurricane activity. The model had a 0.97 correlation between the predicted and observed record from 1949-1990.

Elsner and colleagues began producing multi-seasonal predictions for the Atlantic basin using a mixture of time-series modeling and Bayesian statistical methods. In their first of such efforts, Elsner and Schmertmann [78] proposed using a Poisson distribution (see Appendix **??**) to model annual hurricane counts. Elsner *et al.* [79] attempted to learn and model the de-trended annual TC count time-series, using a univariate autoregressive moving average (ARMA) algorithm on the three dominant components of the TC count time-series. The trend removal approach revealed three modes at the biennial, semi-decadal and sub-decadal time scales. Each of which were hypothesized to be associated with the stratospheric QBO, ENSO, and low frequency changes in Atlantic SST. However relying solely on TC counts (1886-1996), the potential undercounts in the earlier record, and the limited testing set (1992-1996) provided marginal improvements over climatology.

Subsequent works by Elsner and colleagues focused on predicting the number of U.S. landfall hurricanes. Elsner and Jagger [76] proposed a hierarchical Bayesian model for seasonal landfall predictions. The authors modeled annual hurricane counts as a Poisson distribution, which is dependent on the Poisson parameter (λ). The Poisson parameter was in return conditioned on the Cold Tongue Index (CTI; Avg SST anomalies 6N-6S 180-90W) and NOA. This process is hierarchical as first, the stochastic parameter vector β is computed using the observational data (CTI and NOA), then the Poisson parameter is obtained using β, finally the prediction (annual count) is a result of a sample form the Poisson distribution conditioned on λ.

Minor changes have been introduced since [76], Elsner and Jagger [80] used detrended SST observations to predict future SST and subsequently using the SST projections into a hierarchical Bayesian model to predict future hurricane counts. The algorithm has two components: an ARIMA time series model to forecast average hurricane-season Atlantic SST, and a regression model to forecast the annual hurricane counts conditioned on the predicted SST. The algorithm used Monte Carlo sampling to generate predictive samples of SST and samples of the regression coefficients. This way, forecasts are samples of hurricane counts that combine uncertainty in the predictive SST values with uncertainty in the regression model of hurricanes on SST. The prediction were only a moderate improvement from [79] and compared to actual counts. Jagger and Elsner [80] applied a Bayesian model averaging approach to quantify a predictor's impact on predictions (similar to dimensionality reduction, while considering all predictors weighted by their impact). The highest ranked covariates were: September and June sun radiation (SNN), June NAO, July SST, and July-September SOI. The model indicated that US hurricane probability increases with July SST, July SOI, and June SNN and decreases with September SNN. The consensus model was tested on two Hurricane seasons 2007 (one hurricane observed) and 2008 (three hurricanes observed). The model gave the highest probability for 2 ($\sim 22\%$ probability) and 3 ($\sim 20\%$probability) hurricanes respectively.

Chan *et al.* [40] relied on large-scale indices (see Appendix B) instead of observation variables and assumed that given that atmospheric variables “co-occur” with TCs they cannot be used as precursors. Using these indices the authors built a Smooth Multiple Additive Regression Technique (SMART) model that instead of modeling each response as a linear combination of predictors, each response was modeled as a weighted sum of the linear combinations of predictors. The model performed well in hindcast with R=0.88 and RMSE=2.9. Subsequently, Chan *et al.* [55] repeated their forecast a few years later using a longer time-series. Surprisingly, the accuracy in predicting some variables decrease with longer data. This could be due to the change in instrumentation of the later record or potentially due to a change in climatology since the new data was introduced. Additionally, the authors began updating their forecasts in May of the current year causing their forecasts correlation to increase when combine the original and updated models. Similar studies for the Pacific include [54, 53, 97]

NOAA has been issuing seasonal hurricane outlooks for the Atlantic since 1998. Annual forecasts (refereed to as outlooks) are made available to the public in both statistical and deterministic forms, using terciles. The forecasts rely on the state of ENSO, Atlantic SST, and the Tropical multi-decadal mode [45], which incorporates the leading modes of tropical convective rainfall variability occurring on multi-decadal time scales [45]. The seasonal forecasts are not published the peer-reviewed literature. All of their predictions are based on linear regression, yet they do not share their testing and verification methods in their reports.

Gonzalez *et al.* [113] proposed that one of the limitations of previous seasonal prediction models is that they assume the hurricane-atmosphere system to be a single-phase system. Instead, the authors proposed that hurricane activity can be represented as a multi-phase system (low, average, active) and then train a leas-absolute deviation regression model for each phase. To determine which phase the system is in, the authors train a classifier for each feature and the phase is selected by majority voting. Subsequently, each feature has a LAD model associated with each phase of the system. The model produced R = 0.77 and 0.99 RMSE for the Atlantic and North Pacific basins respectively.

In summary, we can see that these models have several limitations. First, given the relatively short record of observational data statistical models are subject to overfilling. Second, the testing methods employed by these models (see Verification Section below) are inconsistent and have limited skill. For instance, a large marjority of these studies compare their predictive ability to that of climatology or the long-term mean for the training data (*i.e.* if the mean number of TC from 1950-1990 was 4 TCs per year, then if a model predicts a storm count closer to the actual observed count than climatology it is considered to be a “skillful” forecast.). More importantly, the interpretability of the model’s output is ambiguous. For instance, if a model predicts a below average season all it takes is a single strong hurricane to inflict major damage therefore rendering the forecast uninformative. This was the case in 1983, when hurricane Alicia struck land during a below average season [79]. Similarly, when one analyzes the performance of a model to predict a composite variable such as NTC, it is unclear which predictors and/or methods are providing skill to the prediction. Finally, as pointed out in [28], these seasonal models have yet to impact climate science. Instead, they are mainly used by the insurance, reinsurance, and tourism industries.

5.4 Short-term forecasts

Short-term forecasts are use mainly by weather services but have received attention in the literature as well. For dynamical models, Belanger *et al.* [17] test the European Center for Medium-Range Weather Forecasts (ECMWF) Monthly Forecast System’s (ECMFS) ability to predict Atlantic TC activity. The model used the TC detection algorithm designed by Vitart *et al.* [247]. For the 2008 and 2009 seasons, the model was able to forecast TCs for a week in advance with skill above climatology for the Gulf of Mexico and the MDR on intraseasonal time scales. The skillful forecast were credited to the model’s ability to simulate the MJO and it ability to capture the large-scale evolution of deep-layer vertical shear, the frequency of easterly waves, and the variance in 850-hPa relative vorticity.

In a similar study, Belanger *et al.* [18] built a forecast model for the North Indian Ocean using the ECMWF Variable Ensemble Prediction System where they forecast genesis as well as pre- and post-genesis tracks. For cyclogenesis, the system had low predictability beyond 48 hours (Brier Skill Score of 0.09 and 0.17 for the Bay of Bengal and Arabian Sea receptively). On average the pre-genesis tracks performed similarly to post-genesis tracks through 120 hours with a total track error growth of 41nm per day.

Another type of short-term forecast models are statistical-dynamical models, such as the National Hurricane Center’s (NHC) Statistical Hurricane Intensity Prediction

Scheme (SHIPS) model [65]. Such models blend both simulations and statistical techniques by producing forecasts based on established historical relationships between storm behavior and atmospheric variables provided by climate models. Such hybrid models, however, have focused mostly on trajectory and intensity predictions [67].

Sencan *et al.* [211] proposed a supervised forecasting of source-sink track dynamics for 10-15 days lead to predict landfall. To do so, the forecasting system learns network constrictions (or motifs) that were associated with landfalling and ocean hurricanes (sinks). Once the sink-based motifs are collected, a feature selection process (decision tree, SVM, and AIC) is used to discriminate between landfalling and ocean hurricanes. Subsequently, given a cyclogenesis region and conditions, the model predicts whether a storm would hit land or not with 90% accuracy.

Hennon and Hobgood [127] (based on work by Perrone for the Pacific basin [194]) combined satellite image analysis with climate variables to classify developing and non-developing storms. They began by manually identifying developing and non-developing cloud clusters from satellite imagery and once a cluster was considered to be favorable for cyclogenesis they recorded the climate variables at that state. The developing cases were further stratified by the number of hours before genesis. A linear discriminant analysis (LDA) was used to separate the cases that developed versus those that didn't develop into tropical depressions. LDA was run 8 times for a 6-48h period (i.e. they classify all clusters that developed within 6h of detection against all non-developing clusters regardless of time of year) The authors found that when $p < 0.7$ genesis rarely occurs, when $p > 0.9$ genesis occurs at 40% of time. The results were composite and case studies. Similar "real-time" works include [194, 128, 210].

In addition to models, climate researchers have relied on data analysis to predict TC activity. Numerous empirical indices have been proposed beginning with Gray's Seasonal Genesis Parameter [114] and Emanuel's Genesis Potential Index [89]. DeMaria *et al.* [66] created a genesis parameter derived from the 5-day mean of vertical wind shear, midlevel moisture, and vertical stability for the tropical Atlantic east of the Lesser Antilles. The Genesis parameter was defined as:

$$GP = \begin{cases} 100 \times S \times I \times M & \text{if } S > 0,\ I > 0, \text{ and } M > 0 \\ 0 & \text{if } S < 0,\ I < 0, \text{ or } M < 0 \end{cases}$$

where the linearly transformed shear S is given by

$$S = \frac{(25 - S')}{40}$$

the linearly transformed instability parameter I is

$$I = \frac{I'}{2.5}$$

and the linearly transformed moisture variable M is given by

$$M = \frac{-26.0 - M'}{5}$$

and S', I', M' are the unscaled shear, instability, and brightness temperature respectively. This genesis parameter explained 50% of the variance of TC activity that formed in this area during 199599.

The necessary large scale factors identified by Gray [119, 117] were compiled largely from composites of upper air soundings and weather station data relative to the location of tropical cyclogenesis for more than 300 development cases and represent the climatology of the regions where TCs form, rather than genesis factors for individual storms. Additionally, it is unclear how useful such indices are for long-term projections as the statistical relations developed based on past and present climate may not apply under future climate warming scenarios [154].

5.5 Testing of Model Predictions

As it is the case with any predictive model, tests are required to evaluate the model's performance. Statistical climate models tend to be tested on two benchmarks: climatology and persistence. Climatology is the mean TC activity (frequency, intensity, *etc.*) of a period generally starting at 1944 or 1970 (when TC data became reliable) and ending the year prior to the first forecast year. Persistence is a measure where the forecast activity for each prediction in a given year was equal to the prior year's observed activity [190]. Predictions of tropical cyclone activity are expressed deterministically (*i.e.* a forecast of either the exact number of tropical cyclones or a specific range of their numbers in a given ocean basin during the peak season) or probabilistically (*i.e.* probabilities of an active season). Table B.1 has a list of skill measures used in forecast studies.

Deterministic Verification

The verification method most familiar to computer scientists is the root mean square error (RMSE), while this method is used in climate science, other similar methods are employed such as the *mean square error skill score (MSE)* defined as:

$$MSE(\theta) = E[(\hat{\theta} - \theta)^2]$$

where $\hat{\theta}$ is estimator and θ is the estimated parameter.

A similar measure is *the mean square skill score (MSSS)*. The MSSS is the Mean Square Error (MSE) of the forecasts compared to the MSE of climatology. The MSE for a forecast at a grid point (or station) may be given by:

$$MSE_j = \frac{1}{n}\sum_{i=1}^{n}(f_{ij} - x_{ij})^2$$

where x and f denote time series of observations and continuous deterministic forecasts. The MSE for climatology is given by:

$$MSE_{cj} = \frac{n-1}{n}S_{xj}^2$$

where

$$S_{xj}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \mu(x_j))^2$$

The Mean Square Skill Score is therefore given as:

$$MSSS_j = 1 - \frac{MSE_j}{MSE_{cj}}$$

For regions over which the MSSS is calculated, it is recommended that an overall MSSS is provided. This is computed as:

$$MSSS = 1 - \frac{\sum_j w_j MSE_j}{\sum_j wjMSE_cj}$$

where the weighting function, w, is equal to $cos(\theta)$, where θ is the latitude of each corresponding gridpoint.

Another common deterministic verification method is the Pearson correlation coefficient (r)

$$r_{X,Y} = \frac{cov(X,\bar{X})}{\sigma_X \sigma_{\bar{X}}} = \frac{\sum_{i=1}^{n}(X_i - \mu_X)(\bar{X}_i - \mu_{\bar{X}})}{\sqrt{\sum_{i=1}^{n}(X_i - \mu_X)^2}\sqrt{\sum_{i=1}^{n}(\bar{X}_i - \mu_{\bar{X})^2}}}$$

where $cov(X, \bar{X})$ is the covariance of the observed and predicted variables and σ is their standard deviation.

or the *Spearman rank correlation coefficient*:

$$r_s = \frac{sum_i(x_i - \mu_X)(\bar{x_i} - \mu_{\bar{X}})}{\sqrt{\sum_i(x_i - \mu_X)^2 \sum_i(\bar{x_i} - \mu_{\bar{X}})}}$$

where x_i and $\bar{x_i}$ are the ranked values of the observed and predicted data (sorted in ascending order by their observed and predicted values). μ is the mean of the actual observed and predicted data.

When the correlation is applied to a small subset of a much longer climatological base period, the *uncentered correlation coefficients* are used instead. Uncentered correlation coefficients are similar to Pearson coefficients, except that terms in the numerator and denominator are not centered on their respective means.

$$r_u = \frac{\sum_{i=1}^{n}(X_i - \mu_X)(\bar{X}_i - \mu_{\bar{X}})}{\sqrt{\sum_{i=1}^{n}(X_i)^2}\sqrt{\sum_{i=1}^{n}(\bar{X}_i)^2}}$$

where X and $\bar{X}$ are the observed and predicted values respectively and μ is the mean.

Another correlation method is the *anomaly correlation coefficient* (ACC). It is used to quantify the spatial correlation between forecast and observed deviations from climatology

$$ACC = \frac{\sum_{i=1}^{N}(f_i - c)(o_i - c)}{\sqrt{\sum_{i=1}^{N}(f_i - c)^2 \sum_{i=1}^{N}(o_i - c)^2}}$$

The *Heidke skill score (HSS)* is a measure of how well a forecast did relative to a randomly selected forecast and is computed:

$$HSS = \frac{(P + TN) - E}{K - E}$$

where P are the positive observations, TN are the true negative observations, E is the expected observation by random chance and is given by $E = \frac{1}{K}[(P + FN)(P + FP) + (TN + FN)(TN + FP)]$, FN is false negative predictions, and FP are false positive. Measures the fraction of correct forecasts after eliminating those forecasts which would be correct due purely to random chance.

In a similar fashion, the *Gerrity skill score (GSS)* uses a scoring matrix S_{ij} (i = 1,...,3), which is a reward/penalty matrix. The score is computed as:

$$GSS = \sum_{i=1}^{j=3} \sum j = 1^{j=3} p_{ij} S_{ij}$$

Where the scoring matrix is given by:

$$S_{ii} = \frac{1}{2}(\sum_{r=1}^{i-1} a_r^{-1} + \sum_{r=i}^{2} a_r)$$

$$S_{ij} = \frac{1}{2}(\sum_{r=1}^{i-1} a_r^{-1} - (j-i) + \sum_{r=i}^{2} a_r); 1 \leq i < 3, i < j \leq 3$$

Where:

$$a_i = \frac{1 - \sum_{r=1}^{i} p_r}{\sum_{r=1}^{i} p_r}$$

Finally, *cross validation* (refereed to as Jackknife in the climate literature) is possibly the most common verification method to computer scientists. To test the usefulness of a prediction equation derived from a sample of limited size, the cross validation is often used to simulate an independent sample. The majority of the studies surveyed here employ the following approach:

1. Build the predictive model based on all except one data point (usually seasonal counts).
2. Predict the number of TCs of the held-out year and calculate the error.
3. Repeat (1) and (2) but now include year 1 and exclude year 2.
4. Repeat (3) until all the years have been excluded once.

Probabilistic Verification

A common verification score for probabilistic forecasts is the *Brier skill score* (BSS), which determines what is the relative skill of the probabilistic forecast over that of climatology, in terms of predicting whether or not an event. To compute BSS we must first define the *Brier Score* :

$$BS = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2$$

Then the *Brier skill score* is given by:

$$BSS = 1 - \frac{BS}{BS_{climatology}}$$

Another measure is the *relative operating characteristic (ROC)*, which plots the forecast's hit and false alarm rate using a set of increasing probability thresholds (for example, 0.05, 0.15, 0.25, etc.) to make the yes/no decision. The area under the ROC curve is frequently used as a score. ROC measures the model's ability to discriminate between positive and negative events.

The *ranked probability skill score (RPSS)* [92] measures how well the model classifies observations compared to climatology. To define RPSS we must first introduce the *Ranked Probability Score (RPS)* which informs how well did the probability forecast predict the category that the observation fell into.

$$RPS = \sum_{k=1}^{N}(P_k - O_k)^2$$

where N is the number of forecast categories, P_k is the cumulative predicted probability of the forecast category up to category k, and O_k is the cumulative observed probability up to category k. Given the above RPS formulation, RPSS is defined as:

$$RPSS = 1 - \frac{RPS_{climatology}}{RPS}$$

where RPS is the predicted ranked probability score and $RPS_{climatology}$ is the reference ranked probability score.

After reviewing the most common verifications methods employed by climate scientists we find that: (1) there is no clear consensus on how forecast models should be evaluated; (2) many verification methods are too simplistic (since they use the very uninformed benchmark of climatology); (3) Even standard methods such as cross validation might not be accurate given that it is assumed that each prediction is considered independent, when in fact many climate phenomena aren't.

5.6 S-ENSO Introduction

Seasonal tropical cyclone (TC) forecasting has become an active field of research [78, 79, 80, 149]. A primary driver of seasonal TC activity is the large-scale conditions over the Atlantic basin [**?**], even if synoptic-scale (*i.e.* African Easterly Waves, *etc.*) and stochastic events cannot be accounted for [153, 90].

One of the well-documented precursors of Atlantic TC activity through large-scale conditions is the El-Niño Southern Oscillation (ENSO): the quasi-periodic cycle of warming and cooling of the near equatorial Pacific sea surface temperatures (SST). ENSO impacts the tropical Atlantic through anomalous atmospheric Walker circulation [250]. As a result, TC activity is suppressed due to atmospheric instability in the form of vertical wind shear [115, 214, 111] or tropospheric warming [229]. Consequently, ENSO's warm phase (El Niño) has been associated with low TC activity and its cold phase (La Niña) with high TC activity.

To capture the ENSO cycle, indices such as NINO1+2 and NINO3.4 (NINO indices thereafter) are constructed by averaging the sea surface temperature (SST) anomalies of fixed regions [234]. However, these fixed-location indices have provided limited insight into seasonal TC activity. Figure 5.1 shows the mean number of June-October TCs as a function of the phase of the NINO3.4 index (Niño (warm), Niña (cold), and neutral). There were 10 Niño years, 11 Niña years, and 12 neutral years between 1979 and 2010. During each phase there was an average of 8, 11.33, and 14.4 TC per year respectively. While there are some notable differences in TC activity based on the phase of the NINO3.4 index, these differences in the mean number of TCs are not statistically significant (see the overlapping error bars in Figure 5.1). Therefore, given the current evidence, we cannot significantly distinguish between the total number of TCs solely based on the phase of NINO indices.

5.7 S-ENSO

Given the limited insight afforded by fixed-location indices, we propose a distance-based ENSO index: spatial ENSO (S-ENSO) that tracks both the intensity and location of warming anomalies in the ENSO region. This is achieved by observing the longitude of the highest SST anomaly within a 10° latitude by 40° longitude region in the tropical

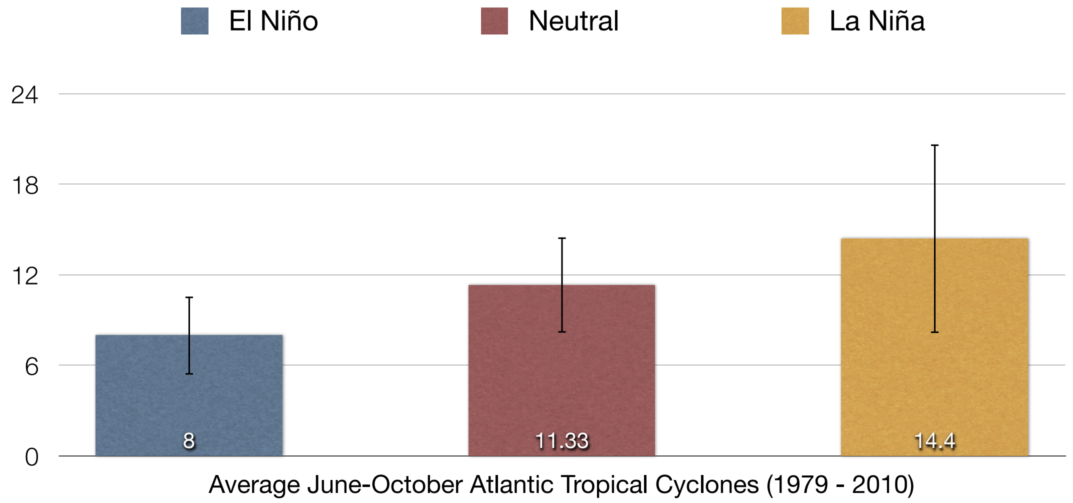


Figure 5.1: The mean June-October Atlantic TC counts for Niño (10 years), Neutral (11 years) and Niña (12 years) years. The ENSO phases are based on NINO3.4 from 1979-2010. Error bars denote one standard deviation for TC counts. While there are notable differences in TC counts based on the phase of ENSO, the large variability of TC counts make discerning ENSO's impact trough traditional indices inconclusive. The NINO3.4 index was calculated using ERSSTv3 monthly anomaly data. The warm and cold ENSO episodes were defined based on the U.S. National Weather Service Climate Prediction Center (CPC) definitions.

Pacific (6° S - 36° N and 90° W - 140° E). Monitoring the spatial distribution of warming has two main advantages over traditional intensity-based indices that monitor a fixed region. First, it inherently accounts for any changes in ENSO's spatial warming patterns [10, 9**?**]. Second, it is intimately related to the east-west propagation of warming along the tropical Pacific associated with changes in the Walker circulation.

El-Niño's long-range impact is through a weakened and shifted Walker circulation. Therefore, it is not only important to monitor the intensity of warming along the equatorial Pacific – something traditional NINO indices do – but also to capture the location of the warming as it informs of the location of updrafts and downdrafts associated with a the Walker circulation. Updrafts over the eastern and central Pacific during El Niño result in downdrafts over the tropical Atlantic and Caribbean and therefore reduced TC activity. In contrast, updrafts over the western Pacific during neutral and La Niña years result in downdrafts over the eastern tropical Pacific and updrafts over the Atlantic, thus leading to more favorable conditions for TC activity. Therefore S-ENSO is better at capturing the east-west SST propagation that is key to ENSO's long-range impact.

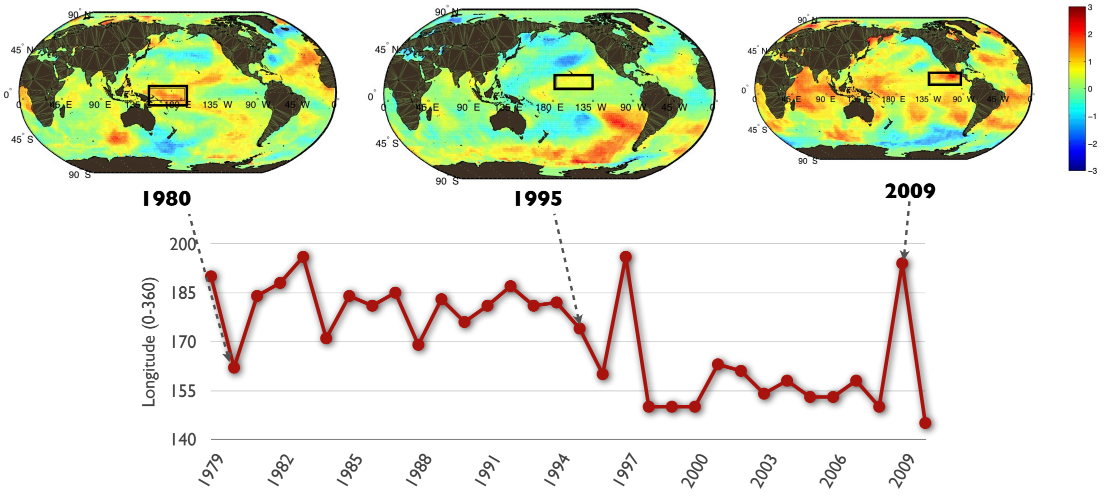


Figure 5.2: An illustrative example demonstrating how the S-ENSO index is built. First, SST anomalies over a certain month range are computed resulting in maps similar to those above. Next, the tropical Pacific is searched for the region with the highest mean SST warming anomaly, represented by the black boxes. Finally, we record the longitude of that region. We repeat this procedure for all years from 1979-2010.

S-ENSO is derived from the ERSSTv3 SST anomaly data [203]. We demonstrate its robustness in predicting Atlantic TC activity using June-October TC counts from the Unisys best track hurricane dataset. We restrict our attention to satellite-era data (1979–2010). Figure 5.2 demonstrates how S-ENSO is calculated. First, SST anomalies over a certain month range are computed. Next, the tropical Pacific is searched for the region with the highest mean SST warming anomaly. This is achieved by sliding a 40 (longitude) × 10 (latitude) window across the tropical Pacific using 1° increments. Finally, we record the longitude of the center that region. We repeat this procedure for all years from 1979-2010. while the approach exhaustively searches for a region with the highest warming anomaly there are no risks of multiple hypothesis testing given that the index is calculated with any knowledge of TC activity.

Figure 5.3 shows the June-October S-ENSO index (A) along with the NINO1.2 (B) and NINO3.4 (C) indices. For the S-ENSO index, a low value (eastward warming) correlates with low TC counts, while a high value (westward warming) correlates with a high count[2].

Figure 5.4 compares how the June-October S-ENSO and the phase of June-October NINO3.4 relate to June-October TC counts. The bar heights denote the value of the S-ENSO index. The color of the bar indicates the phase/strength of the NINO3.4 index. Finally, the numbers topping each bar represent the June-October Atlantic TC counts. Previous research (*e.g.* [119, 111]) suggests that the NINO3.4 index is in its warm phase (dark red) less TCs occur, while when in its cold phase (dark blue) more TCs are observed. In S-ENSO's case, eastward warming (high bars) correlate with low TC counts and westward warming (low bars) with high TC counts. We are interested in

The numbers highlighted in red are the years where NINO3.4 was more accurate at projecting the total TC counts. However, relying on S-ENSO during those years would have led to an overestimation of TC counts. The numbers highlighted in green denote years where the phase of the NINO3.4 index was inconclusive yet the S-ENSO index was in the proper phase. During these years solely relying on a fixed-region NINO index would have resulted in significantly underestimating the number of TCs. The remanning unhighlighted numbers represent years where S-ENSO and NINO both accurately projected June-October TC counts.

[2] The index has been inverted to have a positive correlation

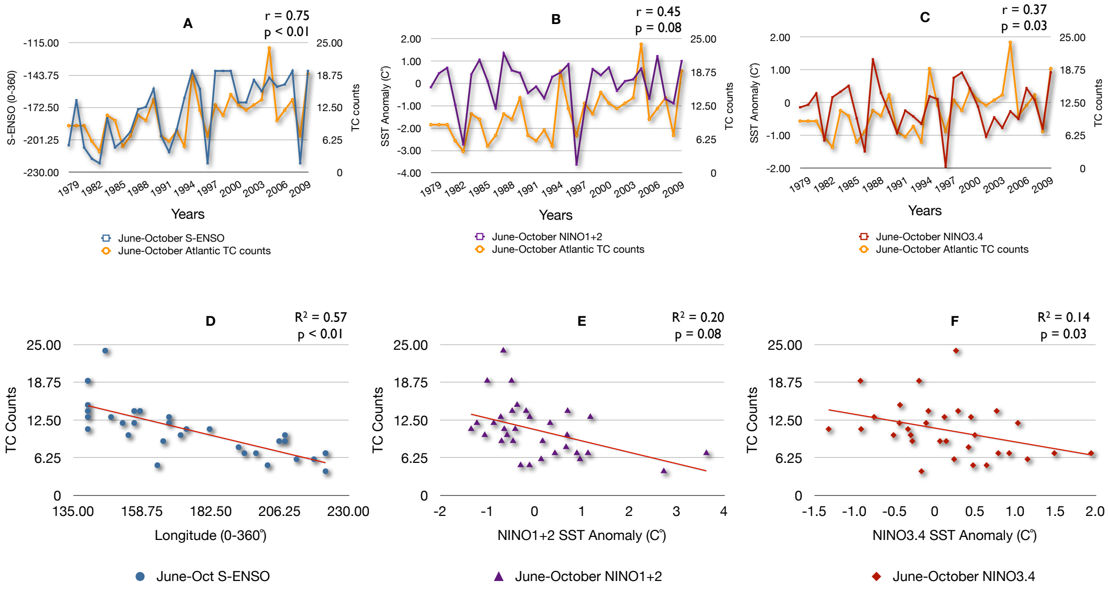


Figure 5.3: The June-October S-ENSO (A), NINO1+2 (B) and NINO3.4 (C) time series are compared to June-October Atlantic TC counts (1979-2010). The y-axis for the indices was inverted to produce a positive correlation. D-F: Same as top but in scatter plot format to show each index' goodness of fit. The p-values are calculated assuming 30 degrees of freedom.

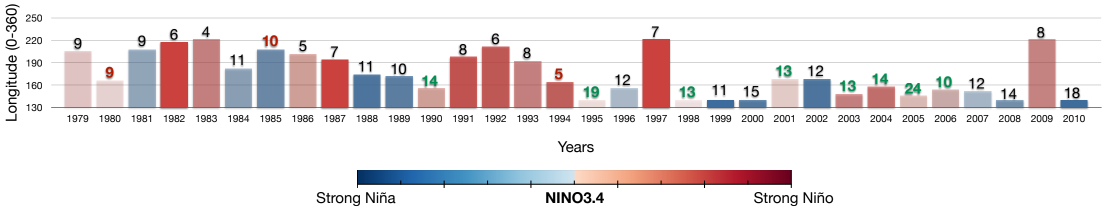
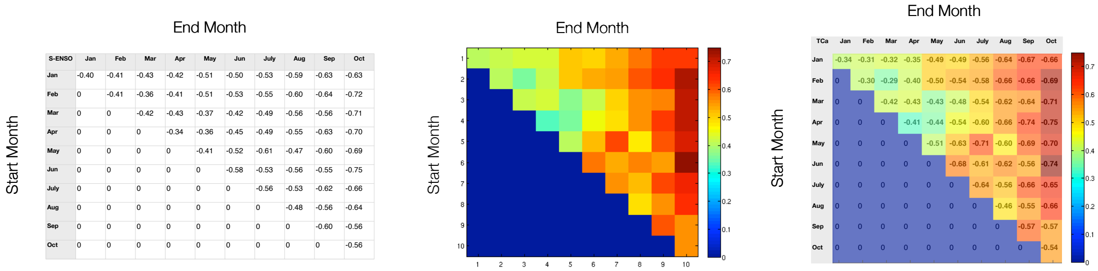


Figure 5.4: Comparing S-ENSO and NINO3.4's correlation with Atlantic TCs. The bar heights denote the value of the June-October S-ENSO index and the bar color denotes the corresponding NINO3.4 index. The number on top of each bar denote the June-October Atlantic TC counts. For the S-ENSO index, a high bar (eastward warming) correlates with low TC counts, while a low bar (westward warming) correlates with a high count. When the NINO3.4 index is in its warm phases (dark red) less TCs occur, while when in its cold phase (dark blue) more TCs are observed. The numbers highlighted in green denote years where the phase of the NINO3.4 index was inconclusive yet the S-ENSO index was in the proper phase. The numbers highlighted in red are years where S-ENSO was less accurate than NINO3.4.

5.8 S-ENSO Results

One major shortcoming of NINO indices is their limited forecasting accuracy before the Northern Hemisphere spring, known as the "spring predictability barrier" [253]. We investigated each index' robustness to both month range selection and increasing lead times. Figure 5.5 shows the performance of S-ENSO as a function of months from which SST anomalies are averaged before building the index [3]. The value recorded in each cell of the tables in Figure 5.5 [4] was obtained by correlating monthly S-ENSO and NINO indices between "Start month" (row) and "End month" (column) to TC counts. As shown in Figure 5.6 in the Supplementary Materials, S-ENSO significantly outperforms all NINO indices regardless of which months are used to build the index.



End Month

Start Month

S-ENSO	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct
Jan	-0.40	-0.41	-0.43	-0.42	-0.51	-0.50	-0.53	-0.59	-0.63	-0.63
Feb	0	-0.41	-0.36	-0.41	-0.51	-0.53	-0.55	-0.60	-0.64	-0.72
Mar	0	0	-0.42	-0.43	-0.37	-0.42	-0.49	-0.56	-0.56	-0.71
Apr	0	0	0	-0.34	-0.36	-0.45	-0.49	-0.55	-0.63	-0.70
May	0	0	0	0	-0.41	-0.52	-0.61	-0.47	-0.60	-0.69
Jun	0	0	0	0	0	-0.58	-0.53	-0.56	-0.55	-0.75
July	0	0	0	0	0	0	-0.56	-0.53	-0.62	-0.66
Aug	0	0	0	0	0	0	0	-0.48	-0.56	-0.64
Sep	0	0	0	0	0	0	0	0	-0.60	-0.56
Oct	0	0	0	0	0	0	0	0	0	-0.56

End Month

Start Month

TCa	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct
Jan	-0.34	-0.31	-0.32	-0.35	-0.49	-0.49	-0.56	-0.64	-0.67	-0.66
Feb	0	-0.30	-0.29	-0.40	-0.50	-0.54	-0.58	-0.66	-0.66	-0.69
Mar	0	0	-0.42	-0.43	-0.43	-0.48	-0.54	-0.62	-0.64	-0.71
Apr	0	0	0	-0.41	-0.44	-0.54	-0.60	-0.66	-0.74	-0.75
May	0	0	0	0	-0.51	-0.63	-0.71	-0.60	-0.69	-0.70
Jun	0	0	0	0	0	-0.68	-0.61	-0.62	-0.56	-0.74
July	0	0	0	0	0	0	-0.64	-0.56	-0.66	-0.65
Aug	0	0	0	0	0	0	0	-0.46	-0.55	-0.66
Sep	0	0	0	0	0	0	0	0	-0.57	-0.57
Oct	0	0	0	0	0	0	0	0	0	-0.54

Figure 5.5: The linear correlation coefficients of every possible S-ENSO between Start Month (rows) and End Month (columns). Left, table containing the correlation coefficients for S-ENSO from "Start Month" (row) to "End Month" (column). Middle, the same information in a color map format. Right, the two representations superimposed. S-ENSO's performance increases as the index includes SST anomalies during the Atlantic TC season (June-October).

Figure 5.7 summarizes the performance of each NINO index as well as S-ENSO for pre-season and in-season forecasts. The right panel in Figure 5.7, shows the linear correlation coefficient between all indices ending in October and June-October TC counts. Each column represents the linear correlation between June-October TCs and each index starting from the month listed on the x-axis and ending in October. The vertical red line indicates the 95% confidence interval assuming 30 degrees of freedom. Figure 5.7's left panel shows the same results as the right panel except for an April end month. None

[3] A similar analysis for all ENSO indices is shown in Figure 5.6

[4] As well as in Figure 5.6

S-ENSO

End Month

Start Month

S-ENSO	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct
Jan	-0.40	-0.41	-0.43	-0.42	-0.51	-0.50	-0.53	-0.59	-0.63	-0.63
Feb	0	-0.41	-0.36	-0.41	-0.51	-0.53	-0.55	-0.60	-0.64	-0.72
Mar	0	0	-0.42	-0.43	-0.37	-0.42	-0.49	-0.56	-0.56	-0.71
Apr	0	0	0	-0.34	-0.36	-0.45	-0.49	-0.55	-0.63	-0.70
May	0	0	0	0	-0.41	-0.52	-0.61	-0.47	-0.60	-0.69
Jun	0	0	0	0	0	-0.58	-0.53	-0.56	-0.55	-0.75
July	0	0	0	0	0	0	-0.56	-0.53	-0.62	-0.66
Aug	0	0	0	0	0	0	0	-0.48	-0.56	-0.64
Sep	0	0	0	0	0	0	0	0	-0.60	-0.56
Oct	0	0	0	0	0	0	0	0	0	-0.56

NINO1.2

End Month

Start Month

NINO 1.2	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct
Jan	0.01	0.05	0.10	0.15	0.19	0.23	0.27	0.30	0.33	0.37
Feb	0	0.12	0.16	0.20	0.23	0.27	0.30	0.33	0.36	0.40
Mar	0	0	0.20	0.22	0.25	0.28	0.31	0.34	0.37	0.41
Apr	0	0	0	0.23	0.26	0.29	0.32	0.35	0.38	0.42
May	0	0	0	0	0.29	0.31	0.34	0.37	0.40	0.44
Jun	0	0	0	0	0	0.34	0.36	0.39	0.42	0.46
July	0	0	0	0	0	0	0.38	0.41	0.44	0.48
Aug	0	0	0	0	0	0	0	0.43	0.46	0.50
Sep	0	0	0	0	0	0	0	0	0.47	0.52
Oct	0	0	0	0	0	0	0	0	0	0.55

NINO3.4

End Month

Start Month

NINO 3.4	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct
Jan	0.04	0.02	0.01	0.02	0.06	0.10	0.14	0.18	0.23	0.28
Feb	0	0.01	0.01	0.05	0.10	0.15	0.19	0.23	0.29	0.32
Mar	0	0	0.03	0.08	0.14	0.20	0.24	0.28	0.33	0.35
Apr	0	0	0	0.14	0.21	0.26	0.29	0.32	0.35	0.37
May	0	0	0	0	0.26	0.31	0.32	0.34	0.36	0.38
Jun	0	0	0	0	0	0.33	0.32	0.33	0.36	0.37
July	0	0	0	0	0	0	0.30	0.32	0.36	0.37
Aug	0	0	0	0	0	0	0	0.33	0.38	0.38
Sep	0	0	0	0	0	0	0	0	0.41	0.40
Oct	0	0	0	0	0	0	0	0	0	0.38

NINO3

End Month

Start Month

NINO 3	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct
Jan	0.00	0.03	0.04	0.06	0.11	0.17	0.20	0.24	0.29	0.34
Feb	0	0.07	0.07	0.09	0.15	0.21	0.25	0.29	0.33	0.37
Mar	0	0	0.06	0.10	0.18	0.25	0.28	0.31	0.36	0.39
Apr	0	0	0	0.15	0.24	0.29	0.32	0.34	0.38	0.40
May	0	0	0	0	0.30	0.34	0.34	0.36	0.39	0.41
Jun	0	0	0	0	0	0.36	0.34	0.35	0.39	0.40
July	0	0	0	0	0	0	0.31	0.34	0.38	0.39
Aug	0	0	0	0	0	0	0	0.35	0.40	0.41
Sep	0	0	0	0	0	0	0	0	0.43	0.42
Oct	0	0	0	0	0	0	0	0	0	0.41

NINO4

End Month

Start Month

NINO 4	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct
Jan	0.18	0.15	0.13	0.12	0.09	0.06	0.03	0.00	0.04	0.08
Feb	0	0.13	0.11	0.09	0.06	0.03	0.00	0.04	0.08	0.13
Mar	0	0	0.09	0.07	0.03	0.01	0.04	0.08	0.13	0.17
Apr	0	0	0	0.05	0.01	0.05	0.09	0.13	0.17	0.21
May	0	0	0	0	0.07	0.10	0.13	0.17	0.21	0.24
Jun	0	0	0	0	0	0.13	0.16	0.20	0.23	0.26
July	0	0	0	0	0	0	0.18	0.22	0.26	0.28
Aug	0	0	0	0	0	0	0	0.25	0.28	0.30
Sep	0	0	0	0	0	0	0	0	0.30	0.31
Oct	0	0	0	0	0	0	0	0	0	0.31

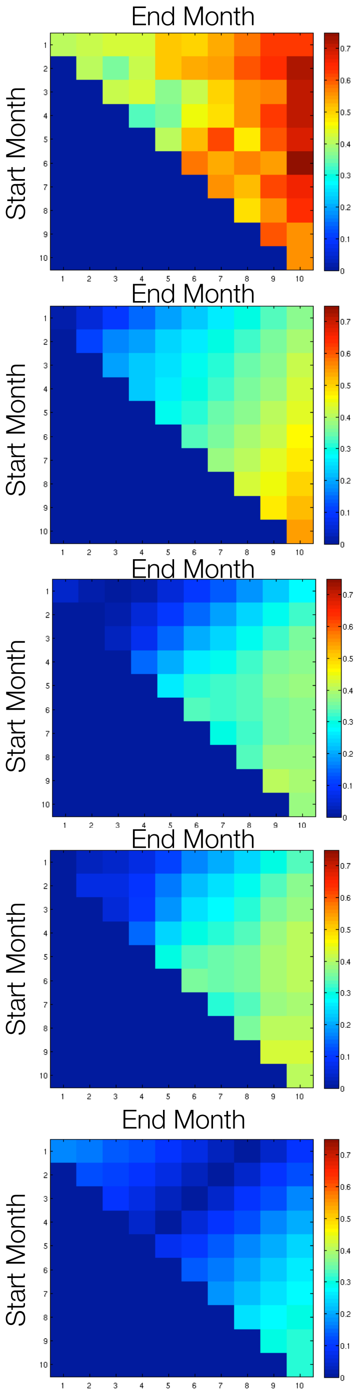


Figure 5.6: The linear correlation coefficients for each index based on which month range was used to average the SST anomalies before building the index. The values in each cell were obtained by first averaging SST monthly anomalies between Start Month (rows) and End Month (columns). Given that selecting a favorable month range might bias findings, here we show that S-ENSO is superior to NINO indices in predicting Atlantic TC activity regardless of which month range one uses to build the index.

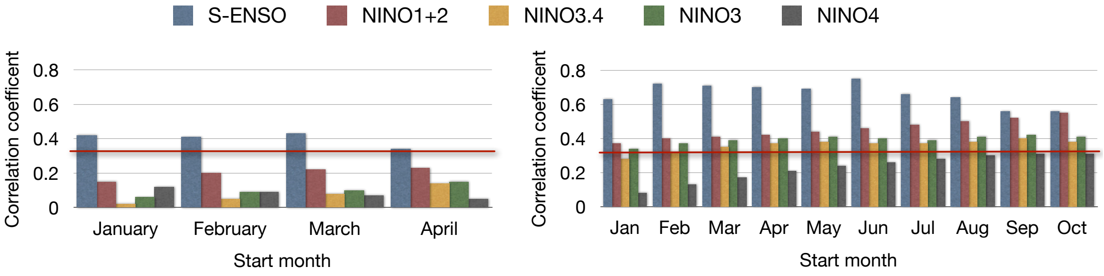


Figure 5.7: The linear correlation coefficients between June-October Atlantic TCs and ENSO indices for an April lead time. Each column represents the performance of the indices build from start month to April (January-April, February-April, March-April, and April-April). The red vertical line denotes 95% confidence assuming 30 degrees of freedom. Only S-ENSO is significantly correlated with Atlantic TCs for April lead time.

of the NINO indices ending in April have significant correlations with June-October TCs, while S-ENSO is significantly more robust than traditional NINO indices to the spring predictability barrier. These results indicate that S-ENSO is more correlated to TC counts than traditional NINO indices, whether predicting using pre-season SST anomalies or correlation with in-season data.

One possible explanation for the lack of predictive accuracy for fixed-region NINO indices compared to S-ENSO, is an increasing number of studies reporting a shift in ENSO's warming patterns [10, 142, 264, 159, 146]. Here, we examine empirically the extent of such a shift. For every month from January 1979 to December 2010, we monitor the longitude of the warmest 10° latitude by 40° longitude region in the Pacific. As it can be seen in Figure 5.8 there has been a distinct westward shift in the longitude of the warmest Pacific region. Such a shift may render monitoring a fixed region less accurate.

5.9 Conclusion

5.9.1 Summary

In a violation to Occam's Razor, statistical forecasting models have grown increasingly complex without significant accuracy gains. We presented a simple and intuitive spatial ENSO index (S-ENSO) that is a better predictor for June-October TC counts than traditional NINO indices and is more resilient to the "spring predictability barrier".

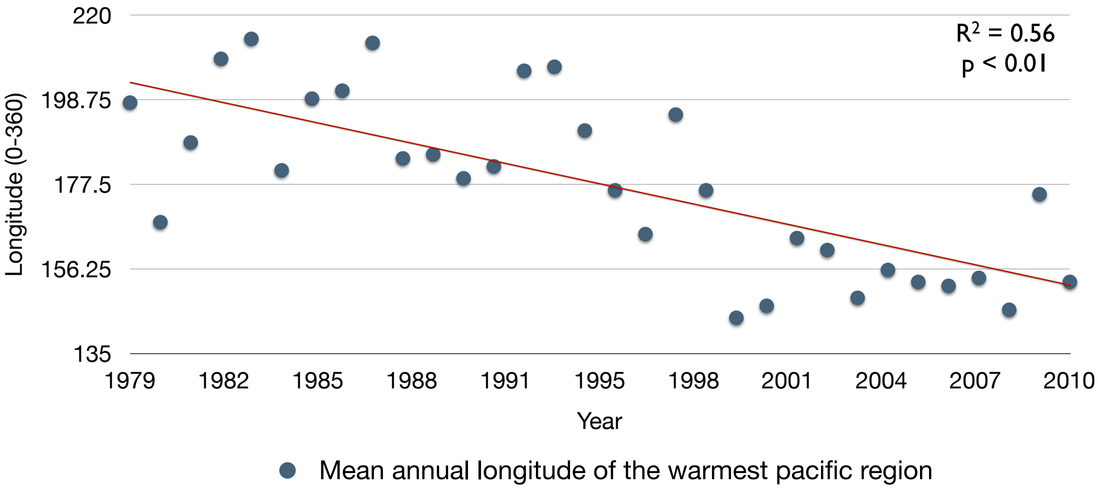


Figure 5.8: The annual mean longitude of the warmest SST anomaly region in the Pacific $(1979 - 2010)$. The figure shows a clear westward shift of the warmest region in the Pacific. $R^2 = 0.54$ $p < 0.01$

Therefore, if dynamical models can resolve the Pacific spatial warming patterns as represented by S-ENSO, they could have significant skill in predicting August-October TC activity.

5.9.2 STDM Contribution

Abstracting the ENSO cycle has been an active field of research. Most attempts have resorted to averaging the SST anomalies of fixed regions, effectively ignoring the spatial component of ENSO such as east-west propagation. As far as we know S-ENSO is the first spatio-temporal attempt at abstracting the ENSO cycle. We demonstrated S-ENSO's ability to forecast and correlate with Atlantic TCs with higher accuracy than traditional NINO indices. We also showed S-ENSO's robustness to search space, region size, and month range. We tested quantities other than longitude such as temperature, pressure, water vapor, etc. None provided significant performances compared to longitude. Finally, S-ENSO provides the foundation to build "evolutionary" statistical models that use the entire spatio-temporal warming patterns to effectively predict future TC activity.

5.9.3 Challenges and Future Work

We have been limited in testing S-ENSO beyond the satellite era. It would be important to extend to earlier periods, although the quality of TC counts is less reliable. One way to address this is to use S-ENSO to predict a proxy variable in the Atlantic that is highly correlated with TC activity and is relatively better represented in pre-satellite era data (*e.g.* MDR SST or Genesis Potential Index). One major limitation of the S-ENSO is that it is dependent on month-range selection. Although we did sensitivity tests to month ranges, the explosion of space-time subspaces between TCs and S-ENSOs makes finding a "parameter-free" relationship difficult. We are currently working on extending S-ENSO to a multivariate index to increase its performance. The major challenge is the multitude of climate variables available, yet only 32 years of reliable data. One way to ensure parsimony is to extend S-ENSO based on the variables that are known to respond to anomalous SST anomalies such as convection.

Chapter 6

Autonomous Ocean Eddies Detection: Spatial Approach

In previous chapters, we highlighted pattern mining as one of the most fundamental data mining applications. In the subsequent chapters we will focus on spatio-temporal pattern mining and show how strategically mining the spatio-temporal context of the data, we can significantly improve the performance of non-space-time algorithms. We demonstrate some of the concepts presented earlier in this thesis using a real-world STDM pattern mining application to identify mesoscale ocean eddies from satellite data. The following chapters provide the reader with concrete examples of challenges faced when mining climate data and how effectively analyzing the data's spatio-temporal context may improve existing methods' accuracy, interpretability, and scalability.

6.1 Introduction

Very much like the atmosphere, our planet's oceans experience their own storms and internal variability. The ocean's kinetic energy is dominated by mesoscale variability: scales of tens to hundreds of kilometers over tens to hundreds of days [262, 204, 46]. Mesoscale variability is generally comprised of linear Rossby waves and as nonlinear ocean eddies (coherent rotating structures much like cyclones in the atmosphere; hereby eddies). Unlike atmospheric storms, eddies are a source of intense physical and biological activity (see Figure 6.1). In contrast to linear Rossby waves, the rotation of nonlinear

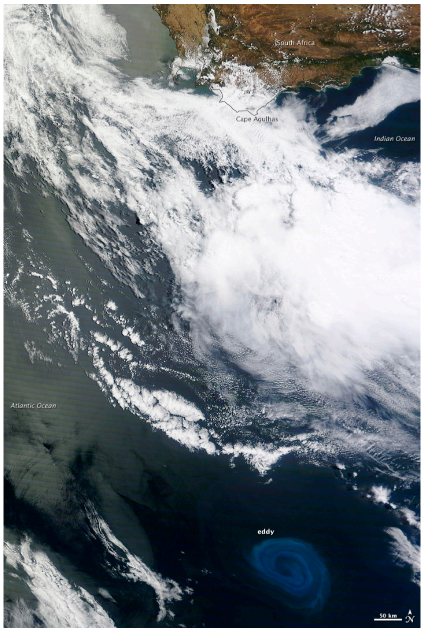


Figure 6.1: Image from the NASA TERRA satellite showing an anti-cyclonic (counter-clockwise in the Southern Hemisphere) eddy that likely peeled off from the Agulhas Current, which flows along the southeastern coast of Africa and around the tip of South Africa. This eddy (roughly 200 km wide) is an example of eddies transporting warm, salty water from the Indian Ocean to the South Atlantic. We are able to see the eddy, which is submerged *under* the surface because of the enhanced phytoplankton activity (reflected in the bright blue color). This anti-cyclonic eddy would cause a depression in subsurface density surfaces in sea surface height (SSH) data. Image courtesy of the NASA Earth Observatory. Best seen in color.

eddies transports momentum, mass, heat, nutrients, as well as salt and other seawater chemical elements, effectively impacting the ocean's circulation, large-scale water distribution, and biology. Therefore, understanding eddy variability and change over time is of critical importance for projected marine biodiversity as well as atmospheric and land phenomena.

Eddies are ubiquitous in both space and time, yet autonomously identifying them is challenging due to the fact that they are not objects moving within the environment, rather they are a distortion (rotation) evolving through a continuous field (see Figure 6.2). To identify and track such features, climate scientists have resorted to mining the spatial or temporal signature eddies have on a variety of ocean variables such as sea surface temperatures (SST) and ocean color. The problem is accentuated further given the lack of base-line data makes any learning algorithms unsupervised. While there exists extensive literature in traditional object tracking algorithms (*e.g.* see Yilmaz et al. [265] for a review), a comprehensive body of work tracking user-defined features in continuous climate data is still lacking despite the exponential increase in the volume of such data [189].

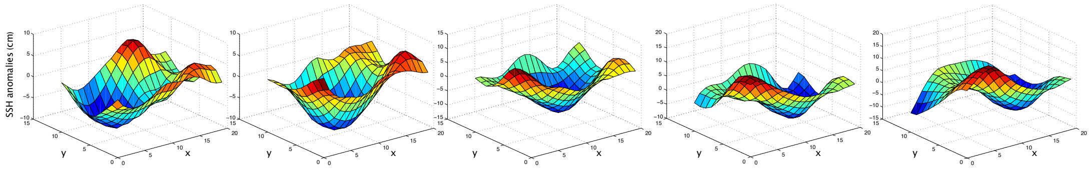


Figure 6.2: An example of a cyclonic eddy traveling through a continuous sea surface height (SSH) field (from left to right). Unlike common feature mining and tracking tasks, features in physical sciences are often not self-defined with unambiguous contours and properties. Instead, they tend to be dynamic user-defined features. In the case of eddies, eddies manifest as a distortion traveling in space and time through the continuous field. A cyclonic eddy manifests as a negative SSH anomaly.

Until recently, ocean eddies were tracked using sea surface temperatures (SST) and ocean color. Now, sea surface height (SSH) observations from satellite radar altimeters have emerged as a better-suited alternative for studying eddy dynamics on a global scale. Eddies are generally classified as either cyclonic if they rotate counter-clockwise (in the Northern Hemisphere) or anticyclonic otherwise. Cyclonic eddies, like the one in Figure 6.3 (bottom panel), cause a decrease in SSH and elevations in subsurface density surfaces. Anti-cyclonic eddies, such as the one depicted in Figure 6.3 (top panel), cause an increase in SSH and depressions in subsurface density surfaces. These characteristics allow us to identify ocean eddies in SSH satellite data. In Figure 6.5, anti-cyclonic eddies can be seen in patches of positive (dark red) SSH anomalies, while cyclonic eddies are reflected in closed contoured negative (dark blue) SSH anomalies.

Figure 6.4 shows two different yet complementary views of eddies and SSH. On the top panel are two anti-cyclonic eddies in the SSH field. The bottom panel shows the temporal profile of a single pixel in the SSH dataset. When taken alone each method has notable limitations. In the spatial view, thresholding the data top-down would force the application to return artificially larger size regions that the eddy occupies (since it

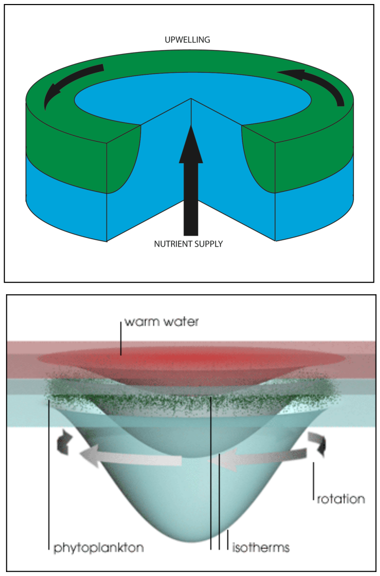


Figure 6.3: **Top:** A schematic cross section of an anti-cyclonic eddy (in the Northern Hemisphere) density surfaces are depressed within the eddy causing an increase in SSH. The elevation of subsurface density surfaces replenishes the upper part of the ocean with nutrients needed for primary production. **Bottom**: A cyclonic eddy causes an decrease in SSH. Bottom image by Robert Simmons of NASA. Best seen in color.

favors the largest region possible). Furthermore, such a thresholding approach is known to merge eddies in close proximity [47]. A temporal view would allow us to identify eddy-like behavior by searching for segments of gradual decrease and increase denoted by the green and red lines [94]. However, a temporal only approach is not enough as multiple pixels must exhibit similar temporal behavior in space and time otherwise the approach would be vulnerable to noise and spurious signals. Our method attempts to combine bother approaches to address each method's limitations. We begin by discussing the spatial approach in more detail.

We present a global eddy monitoring algorithm, EddyScan, that leverages the physical properties of eddies to increase accuracy and scalability compared to existing methods. Our method has three main contributions: first, we do not pre-process the data, effectively increasing the reproducibility of our results. Second, we employ theoretical and empirical findings on global eddy size distributions to reduce the algorithm's computational complexity. Finally, we improve accuracy by separating merged eddies better

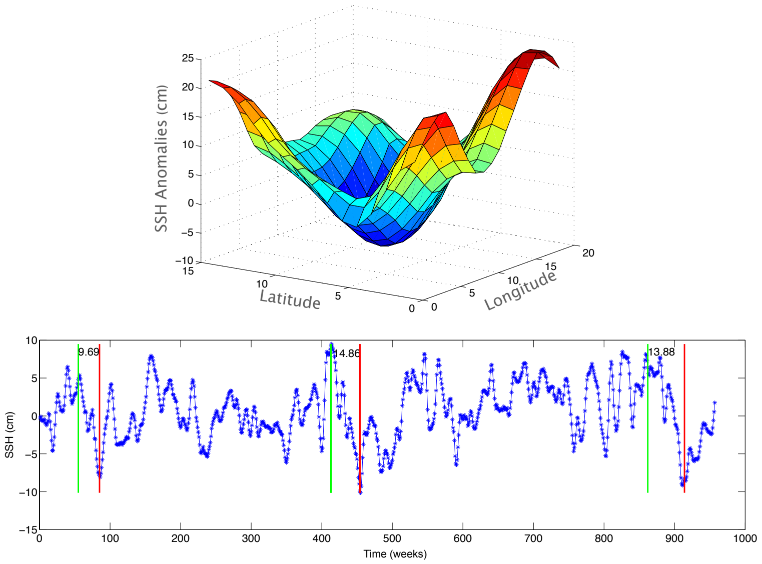


Figure 6.4: Two different but complementary views of eddies' effect on SSH anomalies. *Top*: A three dimensional view of a cyclonic eddy in the SSH field. *Bottom*: an SSH time-series at single location. In both cases, the presence of an eddy is indicated through a sustained SSH depression.

than existing methods.

In the next section, we will briefly review existing eddy tracking algorithms. We then introduce EddyScan and the challenges associated with tracking eddies globally. After that, we present our results and compare them to the eddies identified by Chelton et al. [47] (CH11 hereafter). We conclude the chapter with a discussion of the study's contributions.

6.2 Previous Work

The earliest methods for automatic identification and tracking of ocean eddies relied primarily on proxy variables such as ocean color or SST. The main challenge when using such proxies to track eddies is that they are influenced by a variety of factors in addition to eddies. Thus it is difficult to link changes in those variables to eddy activity alone. Some of the earliest works based on image processing techniques used an edge detection algorithm to detect eddies along the Gulf Stream [133]. Similar image-based

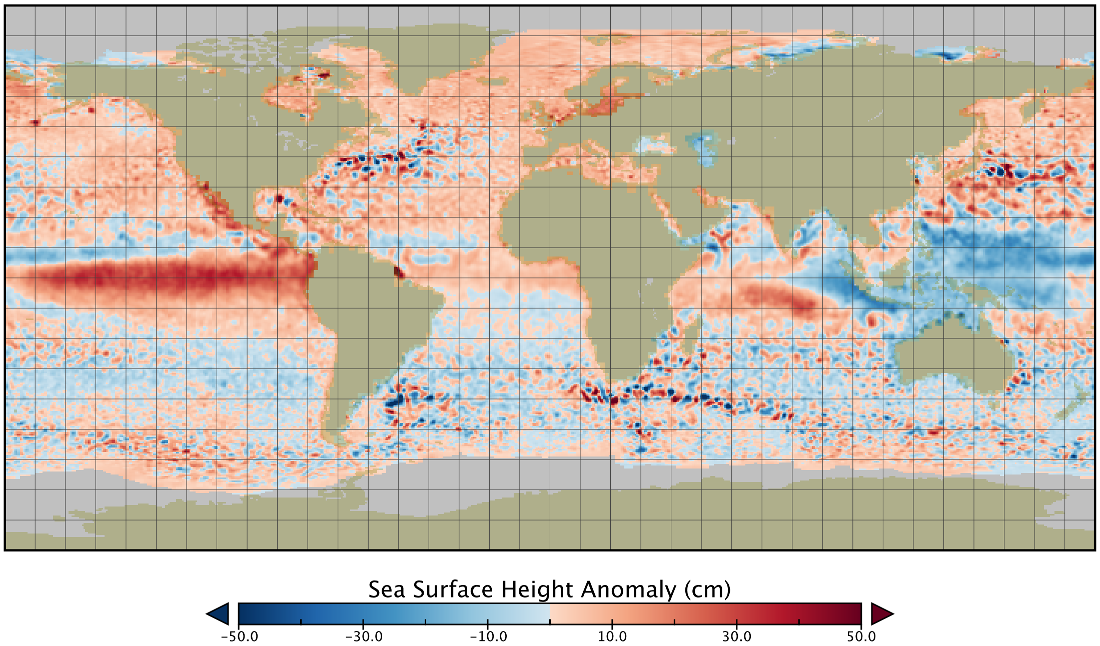


Figure 6.5: Global sea surface height (SSH) anomaly for the week of October 10 1997 from the AVISO dataset. Eddies can be observed globally as closed contoured negative (dark blue; for cyclonic) or positive (dark red; for anti-cyclonic) anomalies. Best seen in color.

algorithms included a neural-network model trained to identify eddies from SST images [34] and an edge detection scheme to isolate eddies between two consecutive SST images [100]. D'Alimonte [64] used the isothermal lines of the SST field to automatically detect eddies. Finally, Dong et al. [72] transformed SST observations into a thermal-wind-velocity field and subsequently tracked eddies in the transformed space.

The recent introduction of SSH satellite observations provided researchers with data that are directly related to ocean eddies. The majority of eddy tracking algorithms define eddies as closed contoured (positive or negative) SSH anomalies (see Figure 6.5). Initial studies built upon techniques developed previously for turbulence simulations [138]. Since then, numerous variations of the approach used by Isern-Fontanet et al. [138] were introduced, *e.g.* [98, 36]. Chelton et al. [46] tracked eddies globally using a unified set of parameters. They also introduced the notion of eddy non-linearity (the ratio of rotational and transitional speeds) to differentiate between eddies and Rossby waves. In the most comprehensive SSH-based eddy tracking study to date, CH11 identified eddies globally as closed contoured smoothed SSH anomalies using a nearest neighbor search. Recently, Faghmous et al. [94] proposed a spatio-temporal approach to monitoring eddies as an alternative to existing image-based approaches. The authors used the eddies' physical properties to sparsify the search space in time and subsequently searched for eddies spatially. A more detailed review of SSH-based eddy detection methods can be found in Appendix B of CH11.

Despite a large body of work, eddy detection algorithms continue to suffer from several limitations. First, water-surface property signatures such as surface temperature or color do not convey much information on the dynamic process of eddies [105]. Second, certain SSH-based methods such as those introduced by Chaigneau et al. [36] use derivatives of the SSH field, which amplify the noise in the SSH signal [47]. Connected component algorithms, such as CH11, tend to be highly parameterized and apply scale-dependent filters to the data to remove features larger (smaller) than a threshold as well as to remove seasonal and internal variability (see Appendix A in CH11 or online supplementary material of [48] for full details on data filtering.) Additionally, CH11 was unable to systematically separate groups of eddies that were merged together because of their close spatial proximity. The spatio-temporal method proposed by Faghmous et al. [94] still had several hard-coded parameters and did not track many eddy properties (such as radius and amplitude).

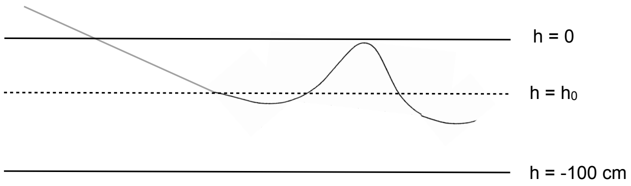


Figure 6.6: Schematic of an anti-cyclonic eddy that is embedded in a large scale background with a larger amplitude than the eddy. If we were to apply a threshold at $h = 0$ the eddy would be missed. This is motivation to use multiple threshold from $h = -100cm$ to $100cm$ as suggested by CH11. Figure adapted from [47]

6.3 Methods

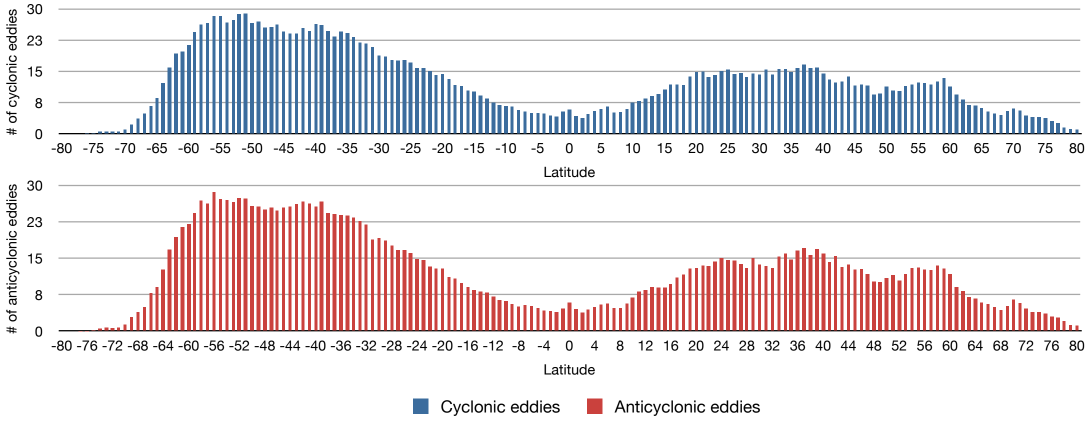


Figure 6.7: Mean weekly cyclonic (top) and anticyclonic (bottom) eddies by latitude as detected by EddyScan for the October 1992 - January 2011 period.

To monitor global eddy activity we used the Version 3 dataset of the Archiving, Validation, and Interpretation of Satellite Oceanographic (AVISO) which contains 7-day averages of SSH on a 0.25° grid from October 1992 through January 2011 [1]. We

[1] Available at `http://www.aviso.oceanobs.com/es/data/products/sea-surface-height-products/`

tracked eddies globally as closed contour of SSH anomalies. This was done in two steps: first we identified features that displayed the spatial properties of an ocean eddy. This was accomplished by assigning binary values to the SSH data based on whether or not a varying threshold was exceeded, and subsequently identifying mesoscale connected component features. We then pruned the identified connected components based on other criteria that are physically consistent with eddies at a given latitude.

Given the large variations in SSH on a global scale, tracking eddies globally presents several challenges: first, SSH data is noisy. Second, until recently, nonlinear eddies were commonly confused with linear Rossby waves in satellite data [48]. Third, eddies can manifest themselves as local minima (maxima) embedded in a large-scale background of negative (positive) anomalies [46] (see Figure 6.6). Therefore, applying a single global threshold would wipe out many relevant features. Fourth, although eddies generally have an ellipse-like shape, the shape's manifestation in gridded SSH data differs based on latitude. This is because of the stretch deformation of projecting spherical coordinates into a two-dimensional plane. As a result, one cannot restrict eddies by shape (*e.g.* circle, ellipse, *etc.*) Finally, eddy sizes vary by latitude, which makes having a global "acceptable" eddy size unfeasible [105].

In section 2.0.2, we discussed some general challenges that arise when mining climate data. Here we briefly review considerations one must take when specifically identifying and tracking eddies on a global scale. First, due to large-scale natural variability in global SSH data (Figure 6.8) complicate the task of finding a universal set of parameters to analyze the data. For example, the mean and standard of the data yield very little insight due to the high spatial and temporal natural variability. Second, unlike traditional data mining where objects are relatively well-defined, SSH data is prone to noise and uncertainty, making it difficult to distinguish between meaningful eddy patterns from spurious events and measurement errors. Third, although eddies generally have an ellipse-like shape, the shape's manifestation in gridded SSH data differs based on latitude. This is because of the stretch deformation of projecting spherical coordinates into a two-dimensional plane. As a result, one cannot restrict eddies by shape (*e.g.* circle, ellipse, *etc.*) Fourth, eddy heights and sizes vary by latitude, which makes having a global "acceptable" eddy size unfeasible [105]. Therefore, applying a single global threshold would wipe out many relevant patterns in the presence of spatial heterogeneity. A more subtle challenges is that eddies can manifest themselves as local minima (maxima) embedded in a large-scale background of negative (positive) anomalies [46] making numerous features unnoticeable. Fifth, eddies can manifest as local minima (maxima) embedded in a large-scale background of negative (positive) anomalies [46] (see Figure 6.6). Finally, other phenomena such as linear Rossby waves or fronts can masquerade as eddy-like features in SSH data [173, 48]. Despite these non-trivial challenges, a more vexing challenge is that the majority of autonomous eddy identification schemes take the four-dimensional feature representation of eddies (latitude, longitude, time, and value where "value" depends on the field) and analyze that data orthogonally in either space or time only effectively introducing additional uncertainty.

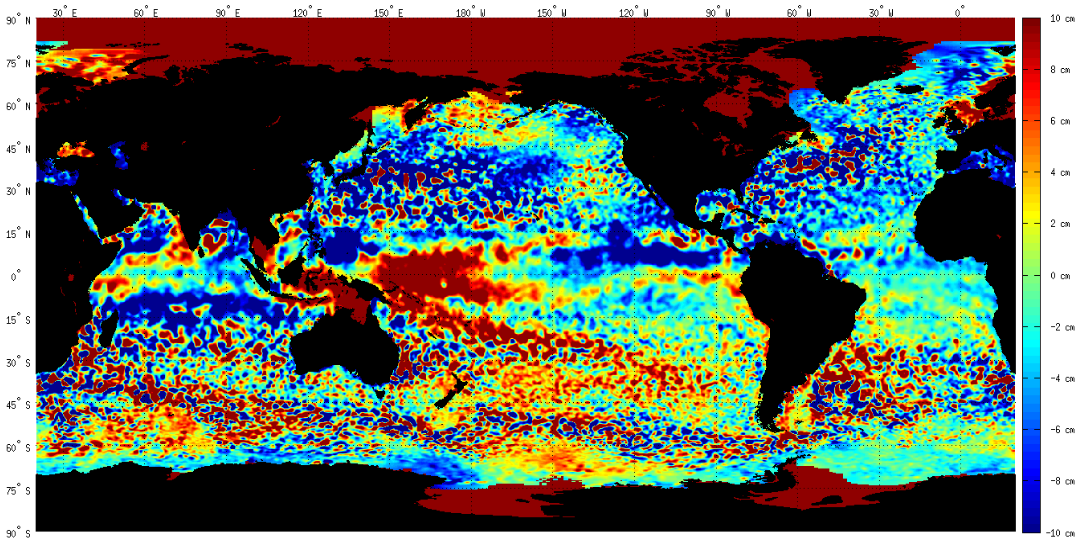


Figure 6.8: Global unfiltered SSH anomalies. The data is characterized with high spatial and temporal variability, where values vary widely from one location to the next, as well as across time for the same location. Therefore traditional measures such as mean and standard deviations yield little insight in global patterns.

Our algorithm addresses some of these challenges, while maintaining physical relevance:

- To address the challenge of identifying mesoscale eddies superimposed on features with larger amplitudes, we repeatedly threshold the data at regular $1cm$ intervals from $-100cm$ to $+100cm$. At each threshold tr_i, we identify all connected components that have an SSH anomaly of at least tr_i. The algorithm then removes from consideration all pixels belonging to the identified connected component and tr_i is incremented. For identification of anticyclonic eddies, tr_i is initialized at $-100cm$ and increased in $1cm$ steps to $+100cm$. Conversely, detection of cyclonic eddies is accomplished by decreasing tr_i from $+100cm$ to $-100cm$. In this way, we identify the largest possible closed contour of an eddy. The gradual thresholding method was proposed by CH11 who also tested sub-centimeter threshold increments but did not observe increased accuracy.

- To address the eddies' varying size by latitude, we use a quadratic function based on theoretical [32, 157, 156] and empirical studies [46, 105, 47] to restrict a reasonable eddy radius based on latitude.

- If an eddy is larger than expected at the latitude (see previous point), there is a chance that two or more eddies were mistakenly merged together. For these "larger than normal" eddies, we apply a convex hull function to determine the size of the smallest convex set that contains all pixels comprising the eddy. If the area of the convex hull is much larger than that of the connected component, it is likely multiple merged eddies and the connected component is not labeled as an eddy. By discarding the connected component, it will remain in the group of pixels to be examined at later thresholds effectively increasing the chance that a higher threshold would eventually break the larger connected component into smaller features.

At a high level, our algorithm extracts candidate connected components from SSH data by gradually thresholding the data and finding connected component features at each threshold. For each connected component, we apply five criteria to determine that it is an eddy: (i) A minimum eddy size of 9 pixels; (ii) a maximum eddy size of 1000 pixels; (iii) a minimum amplitude of 1 cm; (iv) the connected component must contain at least a minimum/maximum and (v) each connected component must have a predefined convex hull ratio as a function of the latitude of the eddy. The first four conditions are similar to those proposed by CH11. The convexity criterion is to ensure that we select the minimal set of points that can form a coherent eddy, and thus avoid mistakenly grouping multiple eddies together. Once the eddies are detected, the pixels representing the eddy are removed from consideration for the next threshold level. Doing so ensures that the algorithm does not over-count eddies. Removing the pixels will not compromise the accuracy of the algorithm given that the first instance an eddy is detected will be at its most likely largest size as a function of the threshold. The main distinction between our implementation, EddyScan, and CH11 are two-fold: First, we use unfiltered data while CH11 pre-process the data. Second, to ensure the selection of compact rotating vortices, CH11 required that the maximum distance between any pairs of points within an eddy interior be less than a specified threshold, while EddyScan uses the convexity criterion to ensure compactness. The primary motivation to use convexity is to reduce the run time complexity of the algorithm from $O(N^2)$ to $O(N)$. An examination of the advantages of using convexity over a predefined maximum distance can be found in the discussion section.

EddyScan's pseudo-code is listed in Algorithm 2. An open-source implementation in MATLAB is available at https://github.com/jfaghm/ClimateCode.git

```
Input: $SSH$
Output: $E$ : global eddy list with corresponding pixels belonging to each eddy;
        $A$ : the amplitude of each eddy; $S$ : the surface area of each eddy
for each timestep $t_i$ do
    for each threshold $tr_i \in \{-100 : 100\}$ do
        if SSH value at pixel $i$ $(p_i) < tr_i$ then
            $p_i = 0$;
        else
            $p_i = 1$;
        end
        Identify all connected component objects left;
        for each connected component $CC_i$ : do
            if $CC_i$ meets criteria listed in text then
                Label all pixels $p_j \in CC_i$ as an eddy;
                remove $p_j$ from data;
            else
                Leave $p_j$ in data;
            end
        end
    end
end
```

Algorithm 2: EddyScan: An automatic global eddy tracking algorithm

6.4 Results

We tracked eddies globally in weekly unfiltered SSH data from October 1992 to January 2011 using the procedure described in the previous section. On a weekly average, there were 2100 cyclonic and 2077 anticyclonic eddies with a larger number of eddies in the Southern Hemisphere. The slight preference for cyclonic eddies is consistent with the findings in CH11[2], although our results are not fully comparable since they only report eddies with lifetimes of at least 4 weeks, while we do not track eddies across timeframes.

[2] Available at: http://cioss.coas.oregonstate.edu/eddies/nc_data.html

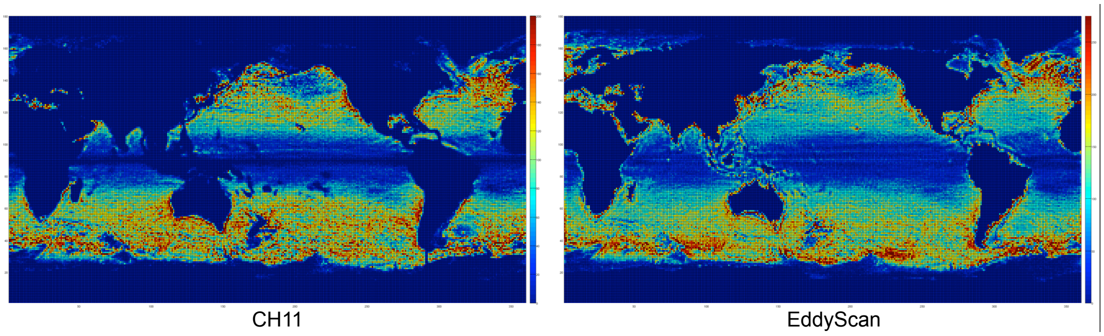


Figure 6.9: Aggregate counts for eddy centroids that were observed through each 1° × 1° region over the October 1992 - January 2011 period as detected by CH11 (left) and EddyScan (right). These results show high eddy activity along the major currents such as the Gulf Stream (North Atlantic) and Kuroshio Current (North Pacific). The high eddy counts along continental and map edges is an artifact of edge effects in the data that we will address in future work. Note the difference in color scale between the the left ([0 − 200]) and right panel ([0 − 270]). This figure emphasizes the similarity in the spatial distribution between the two methods since their exact eddy counts are not comparable (see text). Best seen in color.

Figure 6.7 shows the average latitude-based distribution of cyclonic and anticyclonic eddies. There is a significant difference in the total number of eddies between EddyScan and CH11 (not shown; but they report approximately 3000 eddies per frame to our 4200) most notably at high latitudes and along the equator. The most likely explanation is that the high-pass filtering, while it enhances certain features it also removes others especially at higher latitudes (near the equator) where eddies tend to be small (large). Also CH11's spatial domain spanned 80°N to 80°S while we searched for eddies from 90°N to 90°S.

Figure 6.9 shows the aggregated spatial distribution of eddies on a 1° × 1° grid. Given that our algorithm detects more eddies globally than CH11, due mostly to the fact that CH11 only reports eddies that last four weeks or longer, we used different color scales between the left ([0 − 200]) and right panel ([0 − 270]). The eddy centroid distribution is similar to CH11 (right panel) where high density regions tend to be along currents (*i.e.* Gulf Stream (North Atlantic) and Kuroshio Current (North West Pacific)) and in open oceans.

6.5 Discussion

Our method provides a significant advance to the state of the art by reducing the computational complexity of common connected component algorithms. However, our results depend on a few parameters - most notably data filtering (or lack thereof) and our convexity ratio parameter. Below we analyze the advantages and sensitivity of our results to such parameterazation.

6.5.1 Sensitivity of results to pre-processing

Spatial filtering is a commonly used technique in eddy detection [47]. Numerous studies employ expertly-designed filters to remove signals larger (smaller) than certain scales in addition to filtering seasonality and noise. While filtering has its benefits, the risk of removing important features in the signal is always present. The results presented in the previous section were produced without any preprocessing. To test the sensitivity of the results to filtering, we applied a high-pass filter to remove signals larger than 10° by latitude and 20° by longitude similar to CH11. Figure 6.10 shows the difference in global cyclonic eddies identified by our algorithm using unfiltered (top panel) and high-pass filtered data (bottom panel) for a single time-step. In the filtered data case, we find significantly more eddies at the equator and they tend to be much smaller than expected (eddies have a radius of 200km near the equator [105]). We suspect that these "ghost eddies" are the result of residue noise left after filtering out large features at the equator. Moreover, filtering changes the contours of the data, which means the connected components that result from thresholding a filtered dataset will be geometrically different than those based on the raw data. Because eddy measurements are made on the geometry of the connected component (e.g. surface area) as well as the underlying physical data (amplitude), filtering becomes a source of measurement error.

6.5.2 Advantage of using a convexity metric

As described earlier, binarization of the SSH data during connected component analysis often leads to merging multiple eddies into a single connected component. It is thus necessary to discern those coherent structures representing a single eddy from those comprised of multiple features merged together. Although the vortical character of

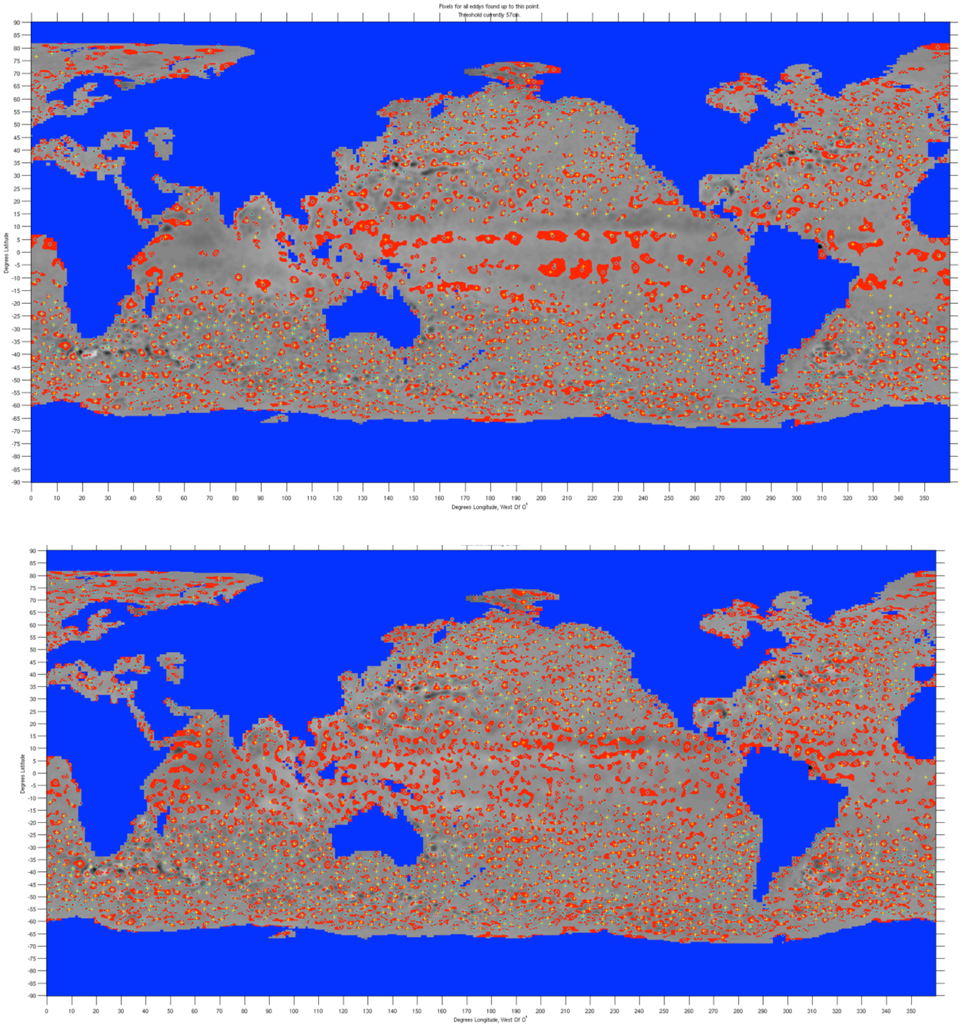


Figure 6.10: An example of the effect high-pass filtering has on EddyScan's output for single time-step. **Top:** Global cyclonic eddies as detected by EddyScan using unfiltered SSH data. The data are on a grayscale for easier visualization of the detected eddies. **Bottom:** Same as top except for high-pass filtered data. High-pass filtering might introduce ghost eddies along the equator as well as noise in eddy characteristics (radius and amplitude).

eddies makes their SSH contours theoretically circular, multiple externalities prohibit us from simply imposing a circularity criterion on connected components. The presence of noise and SSH variability are examples of such factors. In addition, the projection of the SSH data into a two-dimensional grid induces substantial geometric distortions, and re-projecting connected components back onto a spherical surface to analyze their geometry is computationally expensive.

CH11 addresses this problem by restricting any two pixels of a connected component to be within a maximum distance based on latitude. This criterion successfully removes connected components that are particularly eccentric or large, but does not address the issue of merging smaller eddies together (see Appendix B in [47]). More importantly, the maximum distance criterion is extremely inefficient as it requires comparing every pair of pixels within a connected component – a function that grows quadratically with the number of pixels in the connected component. Instead of restricting the maximum distance between any two pixels, we proposed to monitor a feature's convexity to ensure multiple small eddies are not labeled as a single larger eddy.

There are other instances, however, when the maximum distance criterion is unable to avoid merging several smaller eddies together. Figure 6.11 shows an example where the minimal distance between any pair of pixels in the blob is met despite there being several eddies. As a result CH11 (yellow cross) labels the entire feature as a single eddy. EddyScan, however, is able to break the large blob into coherent small eddies.

6.5.3 Sensitivity of results to convexity parameter

Our results depend on the convexity ratio parameter (the ratio of the feature's area to its convex hull area). A ratio of one indicates that the identified feature is perfectly convex. A low ratio indicates a less coherent feature. We tested EddyScan's performance using a variety of convexity ratios and settled on 0.85 because it gave the best balance between cohesiveness and accuracy. Figure 6.12 shows the difference in EddyScan's output based on the choice of convexity ratio. If the convexity ratio is set too low (top panel), large blobs are labeled as eddies throughout the globe dramatically reducing the global eddy count (*e.g.* see lower right corner of the top panel in Figure 6.12). If the convexity ratio is set too high (bottom panel), the global eddy count is not severely affected (the count increased by less than 1% globally) but the mean amplitude and radius are. In

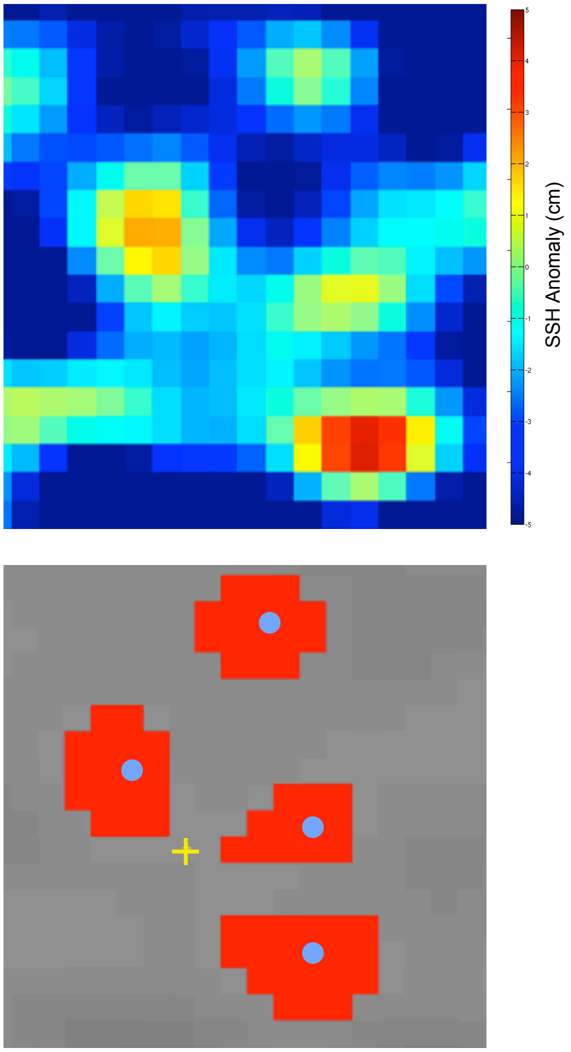


Figure 6.11: An example of when CH11's maximum distance criterion is met, yet the large feature is in fact several eddies merged together. **Top:** a zoomed-in view on SSH anomalies in the Southern Hemisphere showing at least four coherent structures with positive SSH anomalies. **Bottom:** CH11 (yellow cross) identifies a single eddy in the region, while our convexity parameter allows EddyScan to successfully break the larger blob into four smaller eddies. The SSH data are in grayscale to improve visibility of the identified eddies. Best seen in color.

the bottom panel of Figure 6.12 the SSH anomalies are in grayscale for clarity, it easy to see in its attempt to finding the most compact features possible, the contours of many cyclonic eddies are much smaller than expected as shown by the white contours around many eddies (the more accurate labeling would encompass all positive anomalies or white pixels within the eddy's perimeter).

6.5.4 Complexity Analysis

To ensure that only compact rotating vortices are labeled as eddies, CH11 imposes a restriction that the distance between any two pixels in a connected component must be less than a latitude-specific maximum distance. Computing the distance between every pair of pixels in a set of size N would result in a run-time complexity of $O(N^2)$. This operation is prohibitively expensive especially with the additional overhead to convert the distances from pixel/Euclidean to Great Circle distance. In contrast, our convex hull criterion has a runtime of $O(N)$ given that the pixels are already sorted lexicographically (by row) [7], resulting in significant speedup and increased accuracy (by successfully separating merged eddies).

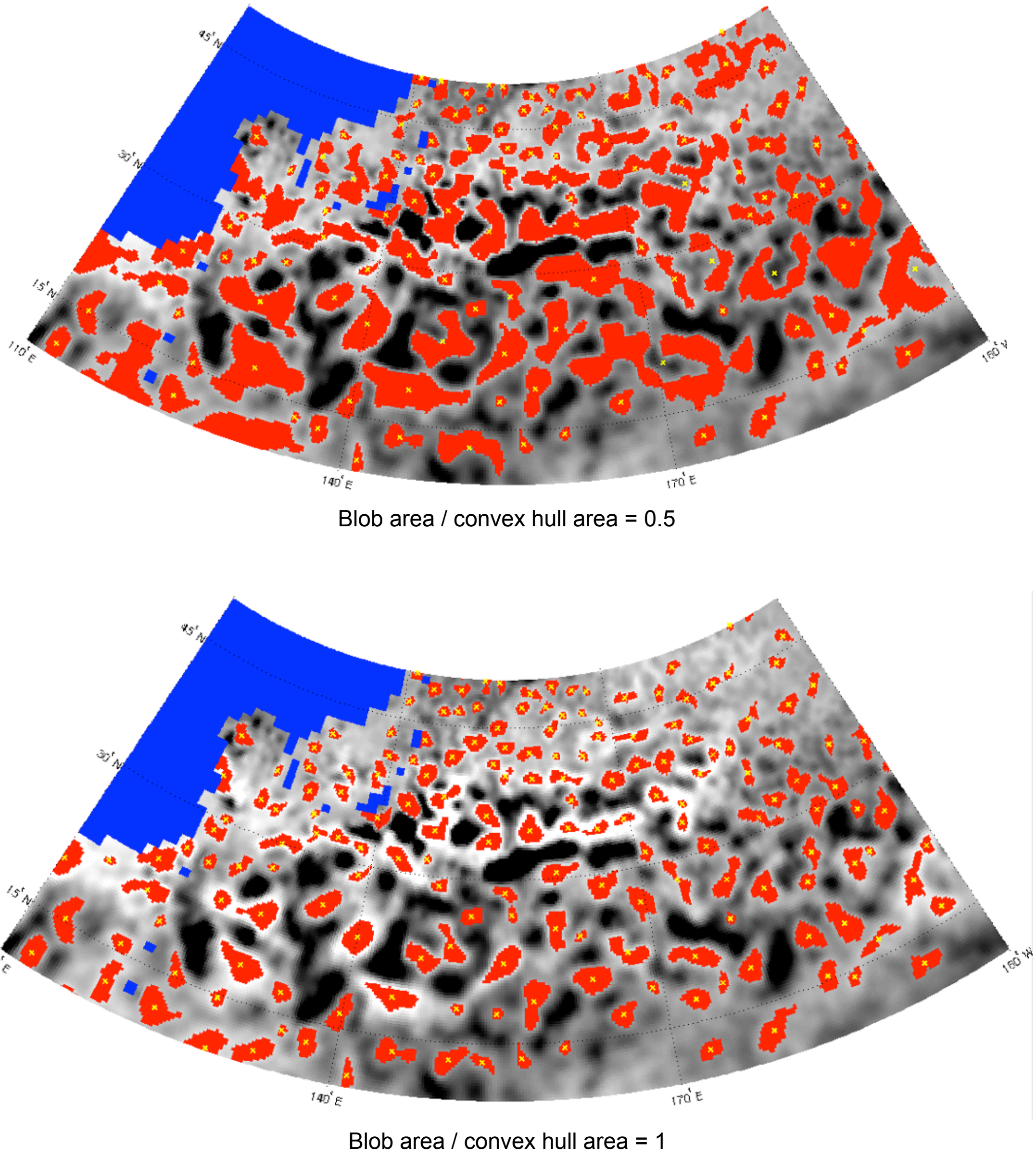
6.6 Conclusion

6.6.1 Summary

In this chapter, we presented EddyScan: an automated, accurate, and scalable eddy detection algorithm from SSH data. EddyScan improves on the state of the art by not preprocessing the data, effectively breaking up merged eddies, and running in a fraction of the time required by traditional eddy detection methods – a significant improvement given the expected dramatic increase in earth science data.

6.6.2 STDM Contribution

This chapter presented one method to mine spatio-temporal patterns in continuous data. This work was inspired by Chelton et al. [47] however we've extend their method to better account for the spatio-temporal context of the data to improve the method's accuracy and scalability. First, we insightfully propose that coherent features such as



Blob area / convex hull area = 0.5

Blob area / convex hull area = 1

Figure 6.12: EddyScan's sensitivity to the choice of convexity parameter. **Top:** When the minimal convexity ratio is set too low (0.5), large incoherent blobs are labeled as eddies significantly affecting the global eddy count. **Bottom:** When the minimum convexity ratio is set to one, identified eddies are much smaller than their actual size because the algorithm picks the most compact features possible. Blue area characterize land. Best seen in color.

eddies should be spatially compact at any given time, despite spatial deformations. To that effect, we imposed a convexity criterion that allows us to maintain compact spatial features over time. Second, we observe that in order to avoid maintaining elongated features, at any time-step the extrema of each feature must be within a maximum distance from each other. This allows us to no longer compare the distance between any two points within a feature as done by Chelton et al. [47].

6.6.3 Challenges and Future Work

Our method suffers from a few limitations: First, it does not track eddies across time, making it susceptible to noise and linear Rossby waves. Second, we do not account for edge cases and coastal regions where over-counting tends to be high. Finally, imposing a minimal eddy size of 9 pixels makes it impossible to identify smaller eddies that are common at higher latitudes. The more vexing issue is that we reduce the information-rich four dimensional data of the SSH field into 2 dimensions (binarizing the SSH) effectively increasing the uncertainty within our application. Furthermore, we do not compute the non-linearity of the features identified and cannot comment on how many features are linear versus non-linear.

Chapter 7

Tracking Features in A Continuos Field: An Ocean Eddy Tracking Application

In the previous chapter, we presented *EddyScan*: a space-based pattern mining eddy identification algorithm. EddyScan improves on the state of the art by not preprocessing the data, effectively breaking up merged eddies, and running in a fraction of the time required by traditional eddy detection methods – a significant improvement given the expected dramatic increase in earth science data. However, left as it was presented, EddyScan would have one serious limitation: it does not track eddies across time, making it susceptible to noise and linear Rossby waves. In this chapter, we introduce the reader to the concept of eddy tracking and present the state of the art. We then, propose a novel eddy tracking algorithm that leverages the spatio-temporal context of the data to improve any mistakes (the unsupervised) EddyScan might have made. We begin with a brief introduction.

7.1 Introduction

Given their importance to ocean dynamics, identifying and tracking eddies has been an active field of research [138, 98, 139, 35, 141, 140, 46, 36, 47, 48, 72, 94, 95]. Traditionally, autonomous eddy monitoring is performed in two independent steps. First, eddy-like features are identified in successive frames of satellite data. Second, the eddy-like features are tracked across time by associating each feature in one frame to another feature in the following time-step. The focus of this chapter is on the tracking phase of this two-step process.

Recently, Chelton et al. [47], hereby CH11, performed the most comprehensive study in autonomous global eddy identification and tracking in sea surface height (SSH) altimeter data. These results were subsequently used in a groundbreaking study published in the journal *Science*, that empirically detailed the impact eddies have on marine biological systems [48]. However, tracking eddies using the most widely used eddy-tracking technique has two major limitations: first it is inexorably dependent on the performance of the automatic eddy identification algorithms. Such methods are highly susceptible to noise and have been known to "miss" features for a few time steps and merge features in close proximity [47]. Second, most eddy tracking methods employ a local nearest neighbor (LNN) tracking algorithm that provides little flexibility for previous errors in detection.

At its core, Eddy tracking is a *data association* problem and lends itself to a family of multi-target tracking applications [37, 122], most notably deferred-logic [196] techniques such as multiple hypothesis tracking [202] that defer making uncertain associations until more data are available. Such methods have been especially useful in cluttered or noisy environments [13]. Extensive work has been done in multi-target tracking and its applications to security [24, 187], computer vision [61], and autonomous agents [184, 181]. We leverage such advances to improve upon the state-of-the-art eddy tracking algorithms, as well as extend multiple hypothesis tracking for applications with noisy features.

This chapter introduces the notion of *multiple hypothesis assignment* (MHA), where a feature may be associated with multiple plausible tracks for a limited time to increase accuracy. Furthermore, we leverage the efficiency of our tracking algorithm to identify and correct any mislabelings that might have occurred during the unsupervised eddy identification phase.

Our method provides several improvements over the state-of-the-art: First, our method is more resilient to noisy observations which is major concern in satellite products. Second, our tracking method is able to maintain tracks even in the event of eddies "disappearing" for a few time-steps and then reappearing due to noise and other sampling constraints. Finally, we employ heuristics to identify and correct possible errors in the eddy identification step in an unsupervised fashion.

7.2 Background and Problem Formulation

Traditionally, the automatic detection and tracking of ocean eddies were achieved using sea surface temperature or ocean color satellite data [193, 99, 72]. The advent of SSH observations from satellite radar altimeters provided researchers with an unprecedented opportunity to study eddy dynamics on a global scale. This is because eddy behavior is intimately linked to SSH. Eddies are generally classified by their rotational direction. Cyclonic eddies rotate counter-clockwise (in the Northern Hemisphere), while anti-cyclonic eddies rotate clockwise. As a result, cyclonic eddies cause a decrease in SSH, while anti-cyclonic eddies cause an increase in SSH. Such impact allows us to identify ocean eddies in SSH satellite data, where cyclonic eddies manifest as closed contoured negative SSH anomalies and anti-cyclonic eddies as positive SSH anomalies.

Eddies are identified in the SSH field by assigning binary values to the SSH data based on whether or not a varying threshold was exceeded, and subsequently saving the eddy-like connected component features that remain after thresholding. Subsequently the identified features are pruned based on physically-consistent criteria that define eddies [47, 95]. Figure 7.1 shows the ubiquitous cyclonic eddy features identified in a single SSH snapshot. Each snapshot contains several thousand eddy features. However that number is often reduced by a variety of significance tests to remove spurious discoveries.

Once eddy features have been identified in all time frames, they need to be tracked across time. Figure 7.2 shows an example of an eddy identified in five successive SSH frames. While tracking a single object is trivial, tracking multiple user-defined features presents unique challenges both conceptually and computationally. First, since the

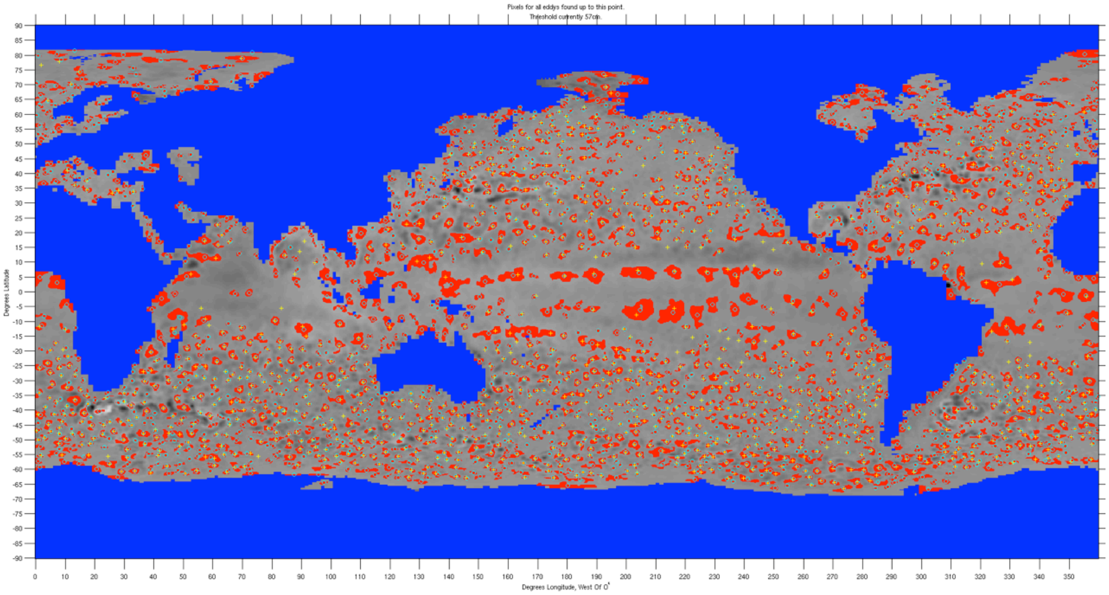


Figure 7.1: Global cyclonic eddy features identified in a single SSH time frame. We identify several thousand eddy-like features in any time frame. The goal is to track eddies by connecting these nearly ubiquitous features with their corresponding features in subsequent time frames.

objects being tracked are not self-defined with clear boundaries like physical objects (*e.g.* car, ball, *etc.*) the very notion of an object is subject to interpretation and errors. This can be seen in Figure 7.2, where the size and shape of the feature changes across frames due to noisy measurements. Second, given the multitude of objects in any given frame, the computational resources required to maintain a large number of potential tracks grows exponentially. Finally, splitting and merging are regular behaviors of ocean eddies that are not common concepts in the traditional object tracking literature.

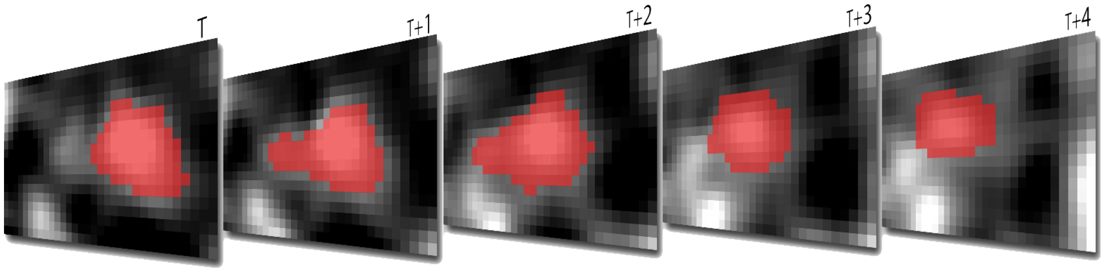


Figure 7.2: A moving eddy as identified in five successive time frames of SSH satellite data. Note how the feature changes size and shape from on frame to the next due to noise in the SSH field.

The majority of eddy tracking algorithms employ a computationally modest, yet limited approach where a feature is connected to the nearest feature in the subsequent time frame. While this approach gives reasonable results in eddy tracking applications [47], its greedy nature means that there will often be cases where it performs sub-optimally, especially in noisy or cluttered environments [13, 24].

Based on the following overview, we proposed the following problem definition and present the methods used to solve such a problem in the following section.

Problem Definition

Given T frames of eddy features identified in an independent detection step. With each frame i having N_i eddy features. Extract globally coherent eddy tracks, where an eddy track is a group of eddy features, satisfying the following constraints: (1) each track has at most one feature from consecutive frames and (2) each feature is associated to at most one track.

7.3 Methods

We implemented *local nearest neighbor* assignment (LNN): the most widely used method in ocean eddy tracking [47]. We also implement *multiple hypothesis assignment* (MHA) adopted from Blackman [24], where each feature is associated with multiple potential tracks. An interactive eddy tracks viewer to compare results between the various implementations is available at: `https://github.com/aaaiDemo`

7.3.1 Local Nearest Neighbor (LNN)

In the LNN scheme, for each eddy feature identified at time t, the features at time $t+1$ are searched to find the closest feature within a pre-defined search space based on the theoretical distance an eddy can travel during the period between two successive time frames. In the event that a feature at time $t+1$ is closest to two features from the previous step, it is assigned to the first encountered feature. This causes the resulting tracks to depend on the scanning order of the feature set.

7.3.2 Multiple Hypothesis Assignment (MHA)

The main difference between MHA and LNN is that unlike LNN, tracks are not formed at every time step. Instead, features are assigned to multiple plausible tracks and uncertain decisions are deferred until more data are available to make a relatively unambiguous track assignment. We use a *track tree* to maintain all possible tracks that be can be constructed starting from any eddy (shown in Figure 7.3). Each path along the track tree is a potential track starting from its root.

In an MHA setting, every track has three outcomes: initiation, expansion, or termination. At any time t, all potential expansions and terminations for the active tracks at $t-1$ are entered in the track tree. Additionally, all features identified at time t initiate potential new tracks by creating new track trees with the new features as the root.

Figure 7.3 demonstrates how MHA maintains all plausible hypotheses across 3 time-steps and 7 eddy features. In this scenario, there are two eddies moving in space and time (e_1,e_3,e_6) and (e_2,e_4,e_7), and a spurious eddy feature e_5. At t_1 the only possibility is to initiate two track trees with e_1 and e_2 as roots. At t_2, three eddy features emerged and they extend the trees created by e_1 and e_2 as well as initiate new trees. Moreover, the tracks created by e_1 and e_2 may end at t_2. Finally two more eddies appear at t_3 and they too are associated with existing trees as well as initiate new ones (denoted by the rightmost single-node trees e_6 and e_7).

While this approach may be more accurate than LNN, the enumeration and maintenance of all possible tracks grows exponentially. To reduce the problem's complexity, we use *gating* and *pruning* functions [202, 160]. Gating ensures that only eddy features that are within a maximum search distance from existing tracks are considered as a potential extension for a track. The gating distance, g, varies based on the location of the track, since the size and distance eddies travel vary uniformly as a function of latitude [105]. An example of gating can be seen in Figure 7.3, where e_5 is not considered as e_1's extension because e_5 is outside e_1's gate.

Furthermore, we employ *N-scan pruning*, where we solve for any ambiguities across all track trees by finalizing the assignment of eddies that appeared during the past N-steps. Hence, every N steps, we reduce the number of paths along each track tree to at most one. To do so, we assign every edge in each track tree a reward $r = g$ - $d(e_i^t, e_j^{t+1})$ if we extend a track at time t such that the distance between the connected features at time t and $t+1$ is $d(e_i^t, e_j^{t+1})$. That is, we give higher scores to tracks that extend through the nearest feature in the search space.

To make final track assignments, we sum the scores of each edge along a path from time $t - N$ to t and select the highest scoring path from each track tree. If two highest scoring paths from different trees share a node (*i.e.* they are ambiguous), then we select the track with the highest total score.

Back in Figure 7.3, If we adopt an 3-scan pruning procedure ($N = 3$), all track trees must be pruned every 4 time steps. In this case, each path along every tree will be scored by summing the rewards r of all edges along the paths between times $t = 1$ $(t - N)$ and $t = 4$. The only tracks remaining from the pruning phase are the highest scoring non-conflicting tracks across all trees. One example of conflicting tracks are (e_1,e_3,e_6), which is the highest scoring track in the leftmost tree, and (e_3, e_6), the highest scoring track in the third tree from the left. They are conflicting because they share e_3 and every feature can only be part of at most one track. In case of conflict, MHA will select the track that has the highest total reward, which in this case is the longest track (e_1,e_3,e_6). Therefore, the tracks (e_1,e_3,e_6) and (e_2,e_4,e_7) are non-conflicting tracks and

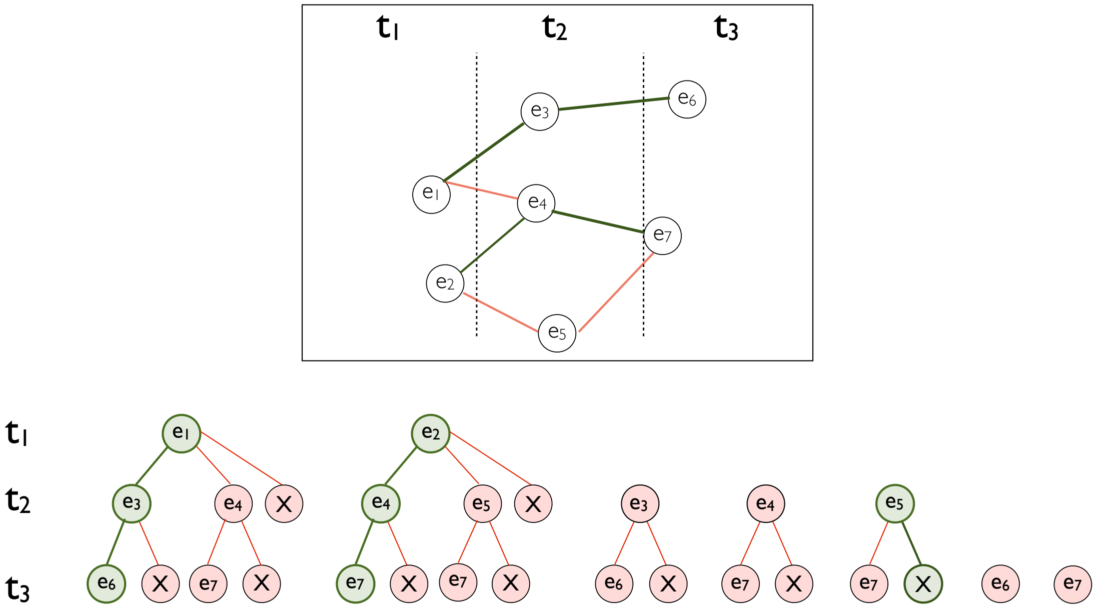


Figure 7.3: An illustrative example of how MHA maintains all plausible tracks in memory. *Top panel:* A two-dimensional spatio-temporal view. Each panel represents one time step. The real tracks are (e_1, e_3, e_6), (e_2, e_4, e_7), and e_5 and are emphasized by green edges. The light red edges are plausible tracks that are considered by MHA but ultimately discarded. *Bottom panel:* the corresponding track trees for all possible tracks based on the data in the top panel. The tracks selected are highlighted in green. Abandoned tracks are in red. "End" nodes are denoted by "X". Note: for simplicity the scoring function, lookahead, and self-learning features are not demonstrated in this example. Figure best seen in color.

together have the globally optimal reward. If new features are identified at $t = 4$ (not shown), then the only possible extensions are for the first and second trees along the highest scored path (green path) or to extend e_6 and e_7. All remaining red paths and trees are pruned and are no longer eligible extension.

Using the same example, LNN is unable to recover the proper tracks in Figure 7.3. At time t_2, LNN will assign e_1 and e_2 to the closest features which are e_4 and since e_4 is already paired with e_1, e_2 is associated with the second closest feature, $e5$. At t_3, LNN will continue to diverge for the proper path, by assigning e_4 to e_7 and terminating e_5. This will result in an incorrect final output of (e_1,e_4,e_7), (e_2,e_5) and (e_3,e_6).

To address the noise in SSH data, MHA extends the traditional multiple-hypothesis tracking algorithm to account for noisy observations, where features contain varying degrees of uncertainty. The two main extensions of MHA that allow it to handle noisy data are: *lookahead*, where we do not terminate tracks that were not extended for a single time step in the event that an eddy "disappeared". The second feature is unsupervised self-learning, where MHA autonomously detects an error in the feature identification phase and takes a corrective measure *a posteriori*.

7.3.3 MHA Lookahead

A common challenge in tracking eddies is that some eddies might be left unassociated from one time-step to the next due to eddies temporarily "disappearing" as a result of noise and sampling errors. CH11 attempted to address this problem by allowing features to not be paired with a subsequent feature for 2-3 time steps in the hope that it will be associated to a track at a later time. However, due to LNN's localized greedy nature the "procedure for tracking temporarily lost eddies were disappointing. The resulting eddy trajectories often jumped from one eddy to another" [47]. Thus, while this lookahead feature is highly desirable, it was abandoned in the most comprehensive study to date, as it was preferable to shorten eddy tracks than to incorrectly associate eddies to incorrect tracks. MHA implements the lookahead feature by assigning a "missing" node to tracks that were not associated with an eddy instead of terminating them. If two consecutive "missing" nodes are within the same path, then it is terminated. MHA avoids jumping tracks because it maintains all plausible future tracks and will likely choose the more consistent track since it has more data available than LNN does.

7.3.4 MHA Unsupervised Self-Learning

One fundamental difference between our method and existing ones is its ability to identify errors from the eddy detection phase and correct them autonomously. The most common type of misidentification is that of artificially merged eddies. Due to the noisy SSH data, features that are in close proximity are susceptible to being labeled as one large eddy [47]. Faghmous et al. [95] addressed such artificial merges by introducing a convexity measure on the features detected to ensure that only the most compact features are selected, as merged features tend to have abnormal shapes. Nonetheless, due to the global nature of the autonomous identification scheme, some clusters of eddies may still be merged as a single large feature. Assume that at time t there are three eddies e_1, e_2, and e_3 with areas a_1, a_2, and a_3 respectively. When a new eddy new_e appears at time $t+1$ it is automatically associated with e_1, e_2, and e_3. If new_e is related to either existing eddy it should be roughly their size. If new_e is a merged eddy, then its area would be significantly larger than *all* existing eddies. Therefore, if MHA detects a new eddy where all of its associated predecessors in the track tree are at least 1/2 of its size, we flag the new eddy as a potential merge and apply the eddy identification algorithm described in [47, 95] to break up the large eddy into smaller ones. If the self-learning procedure returns new eddy features, we then remove all nodes that contained the artificially large eddy from all track trees and extend all active tracks with the new features. This will cause the trees' breadth to increase by $K - 1$ branches where K is the number of newly identified features. The new features will also initiate K new track trees.

The unsupervised self-learning feature is only possible thanks to the multiple hypothesis nature of MHA. There are significant conceptual challenges in extending unsupervised self-learning to an LNN framework. Assume that an eddy feature e_1 is associated to a significantly larger feature e_2 in the following time frame. In order for LNN to label e_2 as an artificial merge with high probability, it must consider all possible feature pairings in the search space to ensure that there is no other feature, e_i that is of the same size as e_2 and could be attached to it. Even then, LNN could not account for the scenario where e_2 is simply a new large eddy that just initiated and will continue during the subsequent time frames.

7.3.5 Computational complexity

Given M, the maximum number of objects in any time frame, and T time-steps, LNN would compare the distance between every other object leading to M^2 comparisons. Therefore, the worst-case runtime complexity of LNN is $O(M^2T)$.

For MHA, let K be the number of hypotheses maintained for consideration at any given time-step. In the worst case, $K = M^T$ where MHA would store all possible tracks, however using N-scan pruning, MHA only stores the N-most hypotheses resulting in $K = M^N$. The lookahead and self-learning features add additional costs to the algorithm. However their running time cannot be explicitly bound but in most reasonable cases it should be no more than $O(M)$. As a result, the worst-case runtime complexity of MHA is $O(M^{N+1}T)$.

7.4 Results

We identified eddies globally in weekly unfiltered SSH data from October 1992 to January 2011 using the procedure described in [47, 95]. We used the Version 3 dataset of the Archiving, Validation, and Interpretation of Satellite Oceanographic (AVISO) which contains 7-day averages of SSH on a 0.25° grid. Given the absence of eddy baseline data or "ground truth", evaluating and comparing methods is a notable challenge. The majority of eddy tracking studies focus on aggregate results such as track lifetimes and distance propagation. An evaluation that is more relevant to our audience would compare the performance of each algorithm at detecting tracks. Ideally, one would use "ground truth" data where the eddy tracks are known in advance and test how well each method recovers such tracks under varying conditions. One way to generate such data would be through a numerical simulation (*i.e.* ocean model) with idealized eddies as ground truth and then gradually add noise. However, such simulations are computationally expensive and require sophisticated physics-based models to simulate eddies and their trajectories. An alternative would be to use filed studies data, where floats are dropped in the ocean and subsequently tracked and eddies are identified when the float rotate while translating (*i.e.* the float is moving along the translating eddy). However, such data make up a small sample size and are not sufficient to significantly differentiate between the two methods. To address these limitations, we designed an experiment to test the relative robustness of each method to noisy measurements; a significant concern in SSH data.

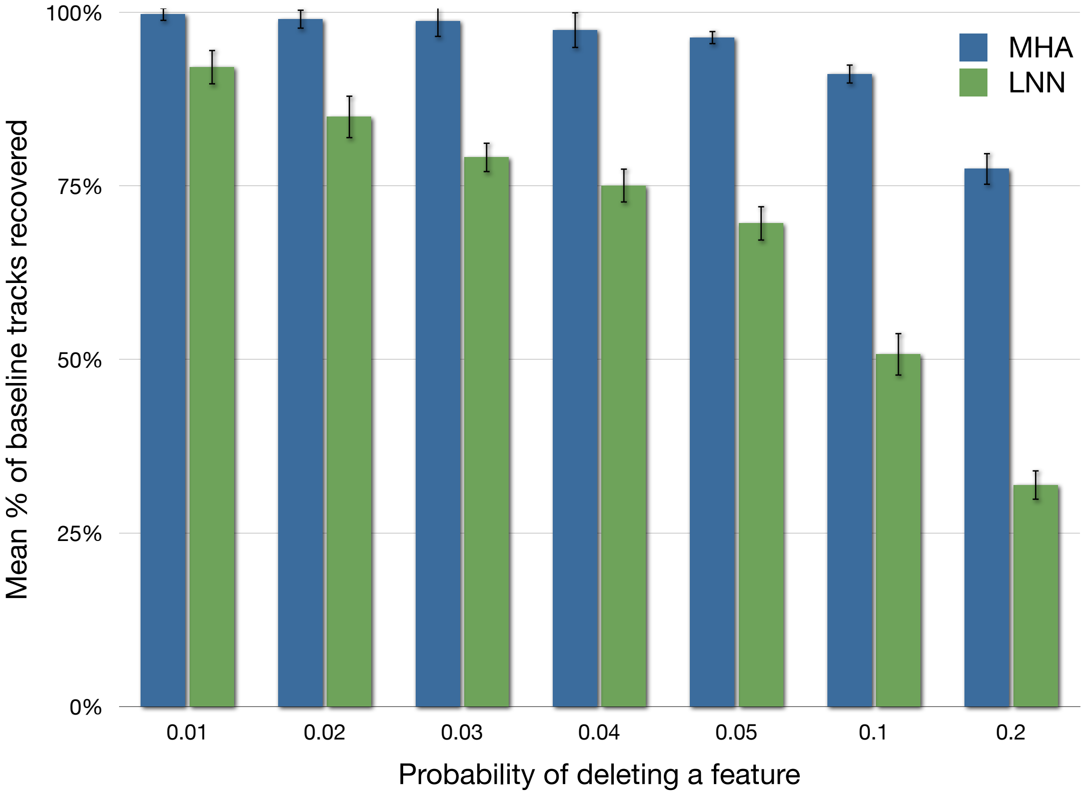


Figure 7.4: The mean percentage of baseline tracks recovered by LNN and MHA as a function of the probability deletion (p). At every timestep, each feature may be removed from the data with a probability p. Each probability level was simulated 50 times and the results show the mean of the 50 simulations. The error bars denote standard deviations.

7.4.1 Sensitivity To Noise

There are three ways that noise can affect automatic eddy monitoring procedures. First, a spurious connected component feature may appear for a single time-step before disappearing. Second, noise may cause uncertain feature boundaries and, as a result, shifted feature centroids. Finally, features might disappear temporarily either due to limitations in the identification scheme, or because of perturbations in the SSH field make the feature unidentifiable. The first type of noise is handled by discarding features that do not persist over four weeks. The second can be handled by smoothing or taking a weighted centroid. The final type of noise is the most vexing. As highlighted earlier, disappearing features is a considerable challenge that CH11 could not address. As a result, we focus our evaluation on the resilience of LNN and MHA to disappearing features.

To test each algorithm's ability to handle missing features, we built a baseline dataset and randomly removed features from that data. We then computed how many of the original baseline tracks were recovered despite the noise. We constructed the baseline dataset by running both LNN and MHA on the features identified in the South Pacific ocean for the years 2005 and 2006. Only tracks that lasted at least 4 weeks and perfectly matched between both methods were selected for inclusion in the baseline dataset. We simulated the event of a feature temporarily disappearing by randomly removing features in the baseline dataset based on a variable "deletion" probability p. To account for a variety of conditions that lead to disappearing eddies, we vary p from 0.01 to 0.2 and run both LNN and MHA algorithms on the noisy dataset. Figure 7.4 shows the results of the experiment. As it can be seen, LNN's recovery rate degrades much faster than that of MHA. It should be noted that a track may have multiple features deleted and that longer-lived tracks are more likely to be missed since there is an increased likelihood of multiple features being removed from longer tracks. However these issues are the same for either procedure, therefore not significantly affecting the results. This experiment empirically shows that MHA is significantly more robust to noise relative to LNN, even in extreme conditions (*e.g.* $10-20\%$ probability of missing a feature).

7.4.2 Impact of Missing Eddies on Track Statistics

LNN's sensitivity to missing eddies can have a significant impact on reported track durations, as eddies disappearing for a single time step would cause tracks to terminate prematurely when the eddy is still moving. Figure 7.5 shows an example where MHA performs better than LNN in the case of an eddy "disappearing" for one time-step. Initially, both methods track the eddy identically. Yet, at time t_3 the eddy identification method is unable to find an eddy and since LNN does not implement lookahead, the track was terminated. MHA, however, was able to recover the full track by allowing the track to be unassociated for a single time-step and then continue tracking once the eddy reappeared.

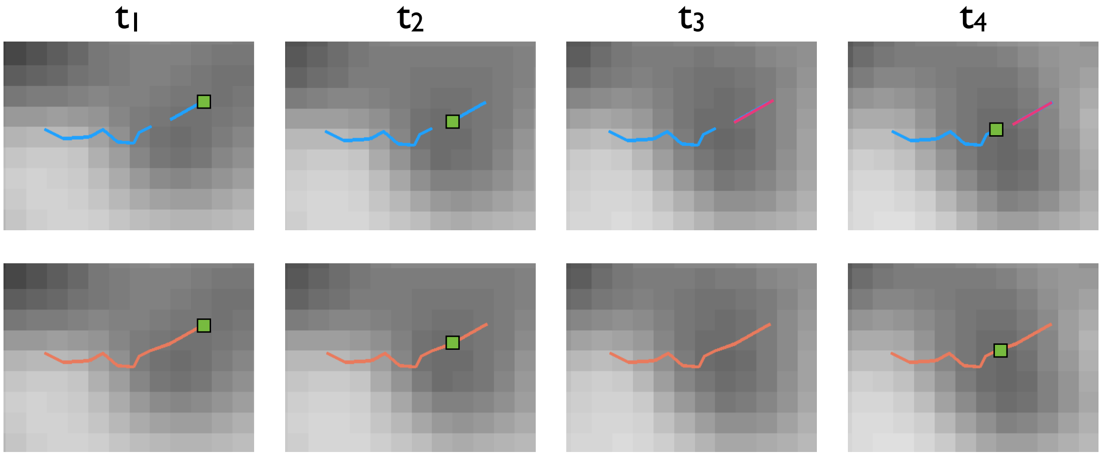


Figure 7.5: An example of how MHA is able to recover tracks even when eddies temporarily "disappear." The eddy is moving westwardly (from right to left). Top (from left): An eddy centroid (green square) and its path as detected by LNN with no lookahead ability. At t_3 the eddy detection algorithm loses the eddy for one time step as it reappears in t_4. However, without the lookahead feature, LNN breaks up the long track into two small tracks as indicated by the different colored tracks. Bottom (from left): the same eddy tracked with MHA. Although the eddy is lost at t_3, MHA successfully recovers the track once the eddy reappears at t_4.

To further quantify the effect of such premature terminations on track statistics, we repeated the noise experiment reported earlier on all features in the South Pacific between 2005 and 2006 (not just the baseline data). Due to space limitations, we only analyze one deletion probability of $p = 0.04$ (the median of the probabilities from the previous experiment). Figure 7.6 shows the mean cumulative track lifetimes for 50 simulations where features have a 4% probability of temporarily disappearing. Although both methods identify nearly the same number of total tracks, LNN identifies 20% less tracks that live 11 weeks or longer. That is a significant difference given that long-lived eddies play a fundamental role in transporting water across ocean basins.

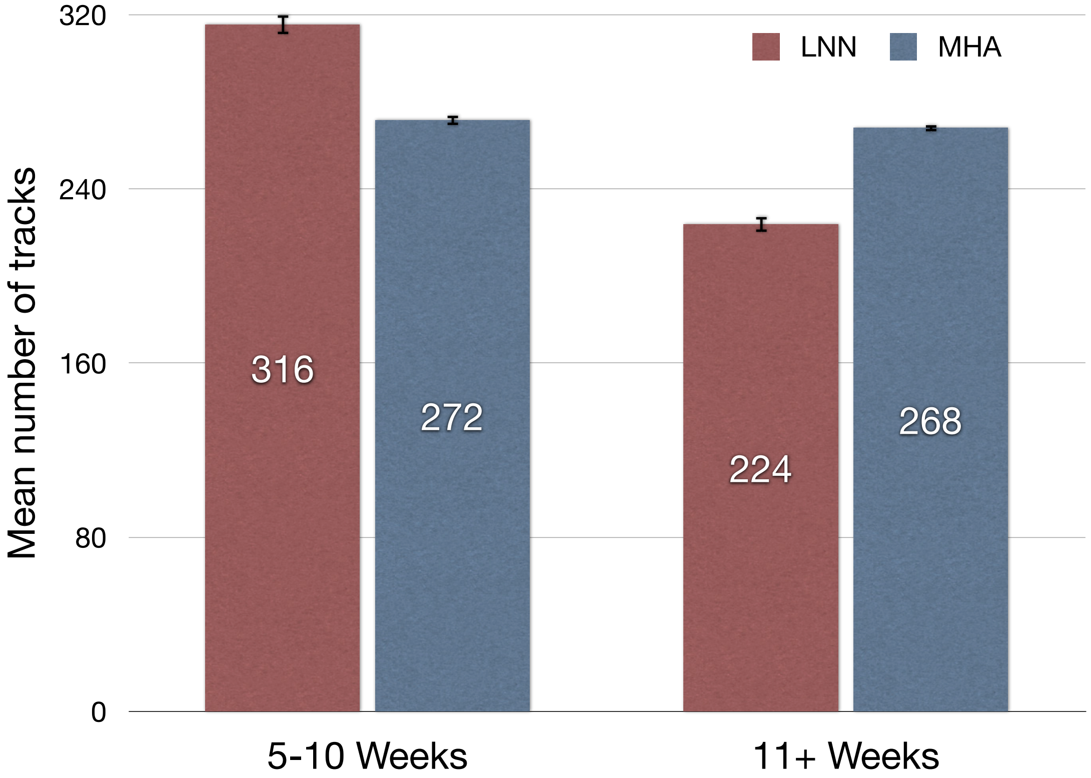


Figure 7.6: Mean cumulative track lifetimes for LNN and MHA tracks from the noise experiment with a probability $p = 0.04$ of a feature temporarily disappearing. MHA identified 20% more long-lived tracks (11+ weeks) than LNN. The simulation was run 50 times and the error bars denote standard deviations.

7.4.3 Impact of Self-Learning on Artificial Merges

Figure 7.7 shows two examples of unsupervised self-learning. In both cases, artificially large eddies were saved in the identification step, but MHA was able to flag such errors autonomously and corrected them by splitting the large features into more accurate smaller eddies. The large panels (A) in Figure 7.7 highlight the artificially merged eddies over a background of greyscale SSH data. These merges were identified by MHA in an unsupervised manner where it would flag any features that had *all* potential tracks associated with it be significantly smaller in previous steps.

The blue ellipses in panel A of Figure 7.7 highlights the merged features in red. Panels B and C show the features as saved in the eddy identification phase in a grayscale and binary forms respectively. The crosses in panel B indicate the local minima, which are centroids of the actual eddies that were erroneously merged. Panels D-E show the resulting eddies once the data correction procedure was performed. In both cases, the separation occurred along the proper direction as denoted by the major axis of the blue ellipses in panel A. These examples demonstrate how the unsupervised self-learning feature effectively leverages the spatio-temporal context of the data to make accurate corrections to the unsupervised eddy detection phase. Such an extension can be used in other multiple hypothesis tracking settings where the objects have varying degrees of uncertainty.

One common sign of an artificial merge is multiple extrema within a single feature. That is because, physically, eddies should only have a single extrema. Using MHA's self-learning feature we were able to autonomously identify 2185 artificial merges within the South Pacific ocean for years 2005 and 2006. Of those potential merges, 1682 features had multiple extrema. An extrema was defined as a pixel whose value was greater/less than all of its 5×5 neighbors. We were able to successfully break 1058 of the reported merges, resulting in a significant increase in feature counts.

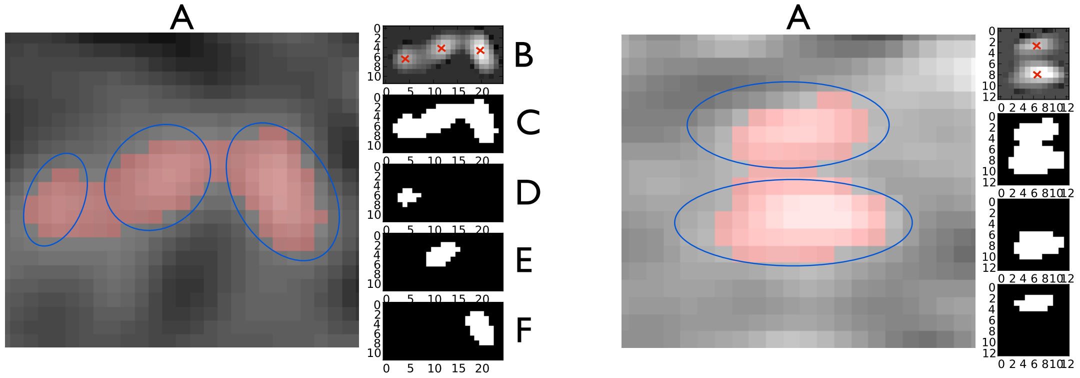


Figure 7.7: Two examples of artificially large eddies being merged as a single eddy and MHA's ability to identify such errors. Panel A denotes the actual SSH data with the large eddy highlighted in red. The smaller vertical panels on the right (B-F) show the binary view of the merged eddy and the corrected data. Panel B shows the actual greyscale SSH data inside the eddy, with the multiple minima marked by red crosses – an indication of an artificial merge. The C panel shows the pixels associated with the eddy as saved by the identification scheme. The bottom two panels (D-F) show the new smaller eddies after data correction.

7.5 Conclusion

7.5.1 Summary

We presented MHA: a multiple hypothesis tracking-inspired eddy monitoring application. MHA extends the traditional multiple-hypothesis tracking approach to handle noisy observations and introduced an unsupervised self-learning feature that refines objects *a posteriori*. Despite the lack of "ground truth" data to evaluate our algorithm, we presented several experiments and case-studies that demonstrated that our method addresses two major shortcomings of the most widely used method in the eddy tracking literature. Furthermore, MHA is an example of how incorporating the data's spatio-temporal context can improve the accuracy of feature identification and tracking algorithms.

7.5.2 STDM Contribution

Our MHA algorithm is able to use its multiple hypothesis nature, along with the data's spatio-temporal context, to take corrective measures on a previously independent eddy identification phase. As far as we are aware, this is the first approach in the eddy tracking literature to do so. Furthermore, our method is more robust to noise as it successfully implements a lookahead feature that makes it less likely to break-up tracks if observations temporarily disappear. All of these features could not be implemented without the deferred logic capability of MHA.

7.5.3 Challenges and Future Work

Additional analysis of MHA's performance based on various factors (*e.g.* eddy density, propagation speed, *etc.*), as well as self-learning's impact on track statistics will be the subject of future work. MHA could also be extended to handle the evolutionary nature of eddies forming and dissipating by adopting MCMC techniques from biological sciences [216]. While MHA's pruning heuristic is based on theoretical eddy dynamics, less rigid heuristics such as the joint probabilistic data association (JPDA) method [14, 13] maybe be more accurate. Finally, now that it has been shown that MHA is a more robust alternative LNN, concepts such as tracks splitting and merging can be incorporated in the MHA scheme by relaxing the constrain of having an eddy be part of a single track at most. Such an addition may provide more accurate information about ocean dynamics.

Chapter 8

Autonomous Ocean Eddies Detection: Temporal Approach

8.1 Introduction

Spatial-based eddy identification schemes often have computational and application-specific limitations. Such algorithms are highly parameterized and rely on complex data-filtering schemes that make reproducibility challenging. More importantly, they fail to capitalize on a critical fact: eddies manifest as coherent SSH distortions in both space and time. When an eddy travels through the SSH field, it leaves a distinctive signature in SSH anomalies in space and time that is wasted when applying a single time-step thresholding method since all features are evaluated in the binary space. Therefore, instead of tracking eddies directly in images of SSH anomalies, an alternative approach could leverage the fundamental spatio-temporal characteristics of eddies.

Eddies form and sustain their energy over a timescale of weeks to months, resulting in gradual changes in SSH on the order of a few centimeters over regions between 50-200 kilometers within the regions where the eddy move. Given the large time-scales within which eddies operate, eddies will manifest as a connected group of gradually increasing/decreasing SSH time-series. We leverage this information to track eddies directly from the SSH time-series as opposed to the SSH heat-maps.

We present an algorithm (adapted from [38]) that monitors the SSH time-series for the unique temporal signal eddies have on SSH. The algorithm operates in three main steps, first we identify individual time-series that have the previously described "eddy-like" behavior. Each candidate time-series will be labeled with a start and end time (t_s and t_e respectively) where a significant gradual increase/decrease occurred. Second, given that an eddy must operate over a large enough region, for each time step t we scan the neighbors of any candidate time-series (where $t_s \leq t \leq t_e$); if a sufficient number of neighbors are also candidate time-series at time t then the identified group is labeled as an eddy. Finally, as the eddy moves from one time-step to the next, we keep adding new candidate time-series as their t_s is reached and remove other time-series as their t_e is passed. We count the duration of each eddy as the number of weeks the minimum number of clustered candidate time-series is met.

Designing intelligent, reproducible and scalable algorithms is crucial for high impact research in oceanography and related fields, and this work is a first step in that direction. In the next section, we will briefly review existing eddy tracking algorithms. In the following section, we will introduce our change detection algorithm, *persistent delta* (PDELTA), which leverages the unique spatio-temporal characteristic of ocean eddies. After that, we will present our results and compare them to the eddies identified by Chelton et al. [47] (CH11 thereafter). We will also compare PDELTA's scalability to that of a connected component algorithm that is similar to the one used by CH11. We conclude the paper with a discussion of the study's contributions and future research directions.

8.2 Methods

Instead of tracking eddies directly in images of SSH anomalies (such as in Figure 8.1), we propose a novel approach that leverages the fundamental spatio-temporal characteristics of eddies. Eddies form and sustain their energy over a timescale of weeks to months, resulting in gradual changes in SSH on the order of a few centimeters over regions between 50-200 kilometers. Given the large time-scales within which eddies operate, eddies will manifest as a connected group of gradually increasing/decreasing SSH time-series. We leverage this information to track eddies directly from the SSH time-series (see Figure 8.2) as opposed to the SSH heat-maps.

Our algorithm (adapted from [38]) operates in three main steps, first we identify

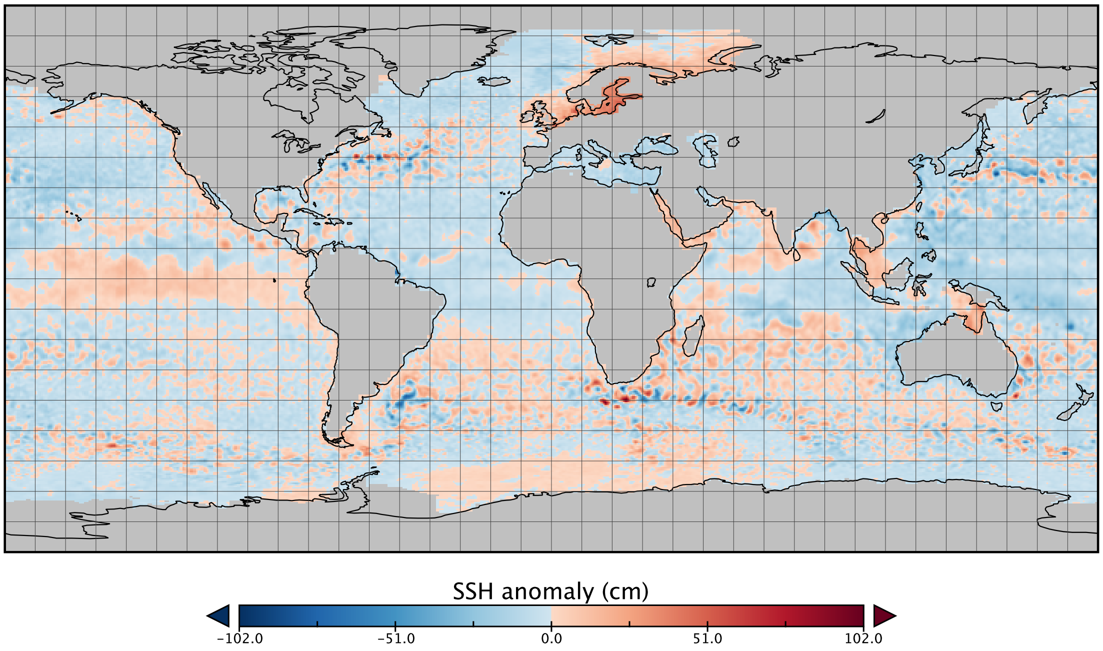


Figure 8.1: Global sea surface height (SSH) anomaly for the week of May 5 1993 from the AVISO dataset. Eddies can be observed globally as closed contoured negative (dark blue; for cyclonic) or positive (dark red; for anti-cyclonic) anomalies.

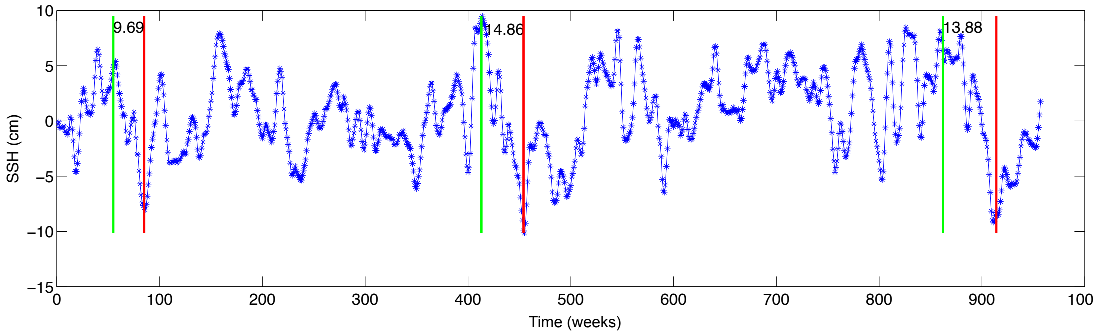


Figure 8.2: A sample time-series analyzed by PDELTA with gradually decreasing segments enclosed between each pair of green and red lines. These segments were obtained after discarding segments of very short length or insignificant drop that are atypical signatures of an eddy.

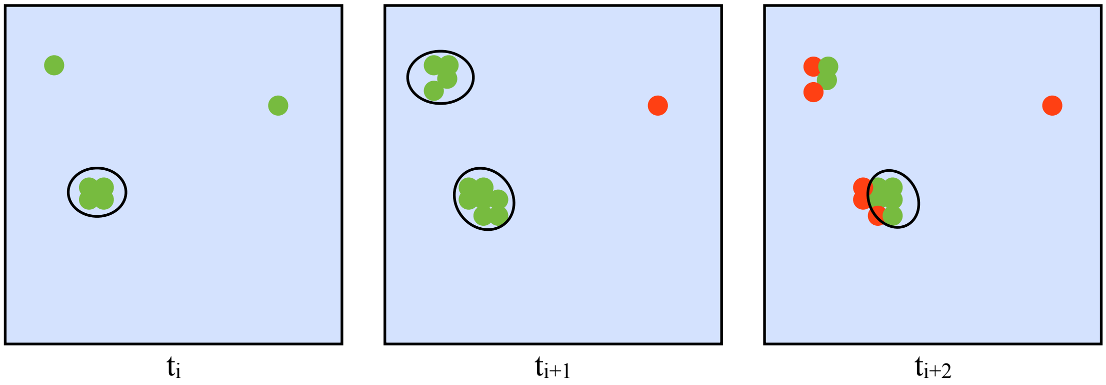


Figure 8.3: An illustration to show PDELTA's spatial analysis component. At any given time t_i only a subset of all time-series are labeled as candidates for being part of an eddy (green points). Only when a sufficient number of similarly behaving neighbors are detected (in this case four) PDELTA labels them as an eddy (black circle). As time passes, some time-series are removed from the eddy (red points) as they are no longer exhibiting a gradual change; while others are added. If the number of similarly behaving time-series falls below (above) the minimum (maximum) number of required time-series, the cluster is no longer an eddy (*e.g.* top left corner at t_{i+2} frame).

individual time-series that have the previously described "eddy-like" behavior. Each candidate time-series will be labeled with a start and end time (t_s and t_e respectively) where a significant gradual increase/decrease occurred. Second, given that an eddy must operate over a large enough region, for each time step t we scan the neighbors of any candidate time-series (where $t_s \leq t \leq t_e$); if a sufficient number of neighbors are also candidate time-series at time t then the identified group is labeled as an eddy. Finally, as the eddy moves from one time-step to the next, we keep adding new candidate time-series as their t_s is reached and remove other time-series as their t_e is passed. We count the duration of each eddy as the number of weeks the minimum number of clustered candidate time-series is met.

Our approach differs from the original PDELTA described in [38] in two main aspects: first, Chamber et al. [38] used PDELTA as a time-series change algorithm only. In our case, we augment PDELTA by adding a spatial analysis feature to properly identify clusters of time-series exhibiting similar behavior. Second, the original PDELTA only detected a single increasing (decreasing) segment in a given time-series. In this variation, PDELTA identifies multiple gradually changing segments within a time-series.

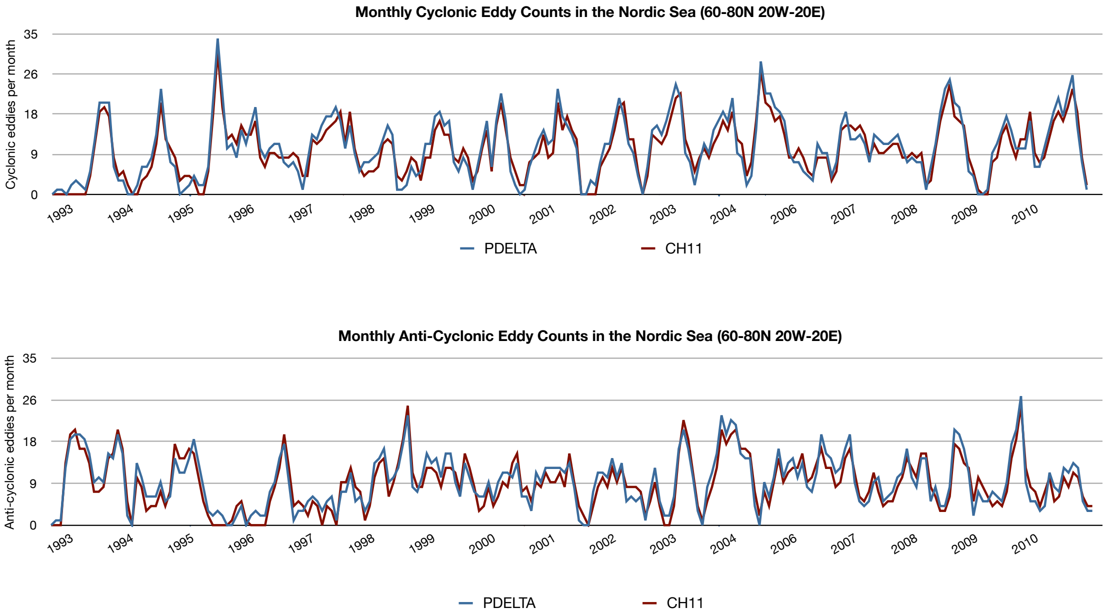


Figure 8.4: Monthly eddy counts (lifetime $\geq$ 16 weeks). **Top:** Monthly counts for cyclonic eddies as detected by our automated algorithm PDELTA (blue) and CH11 (red). **Bottom:** Monthly counts for anti-cyclonic eddies as detected by our automated algorithm PDELTA (blue) and CH11 (red). Overall, PDELTA detected slightly more eddies than CH11. This could be due to an improved ability to track smaller eddies (see text).

This is a useful improvement given that an eddy may appear multiple times at the same location. Furthermore, this feature effectively improves our computational performance by monitoring each location (time-series) at most once. For a complete discussion of the original PDELTA algorithm, including experiments, please refer to [38].

Figure 8.2 demonstrates how our approach detects candidate time-series. The figure shows the SSH anomaly time-series for one grid point in the Nordic Sea. For this particular location, PDELTA identified three segments where a significant gradual decrease in SSH occurred over a long time period starting at approximatively weeks 60, 410, and 870 respectively. During each decreasing segment, we search this location's neighborhood for time-series with similar gradual decrease. Once the significant decreasing segment ends, either there will be other neighbors that will continue to form a coherent eddy or the eddy has dissipated if the minimum eddy size is no longer met (see Figure 8.3).

A spatio-temporal approach to eddy detection and monitoring has several advantages: First, we don't apply any filters to our data and have a minimal number of parameters relative to existing approaches. Second, given that we only consider time-series that experience gradual changes in SSH, our algorithm's search space is significantly reduced compared to searching every pixel's neighbors. Finally, since we incorporate a spatio-temporal detection mechanism we can relax some of the minimum/maximum eddy size requirements that space-only connected component algorithms have.

8.3 Results

We tracked eddies weekly from 1992-2011 in the Nordic Sea region ($60 - 80°$ N and $20°$ W $-$ $20°$ E) and compared the results with those of CH11[1]. We used the Version 3 dataset of the Archiving, Validation, and Interpretation of Satellite Oceanographic (AVISO) which contains 7-day averages of SSH on a 0.25° grid from October 1992 through January 2011.

PDELTA detected slightly more cyclonic (9.89 per month) than anti-cyclonic (9.48 per month) eddies. These differences are consistent with the findings of CH11. Overall, we identified a total of 9.08 eddies per month versus 8.87 for CH11. This could be due to the fact that eddies tend to be smaller in the region analyzed, and thus could have been

[1] Available from: `http://cioss.coas.oregonstate.edu/eddies/nc_data.html`

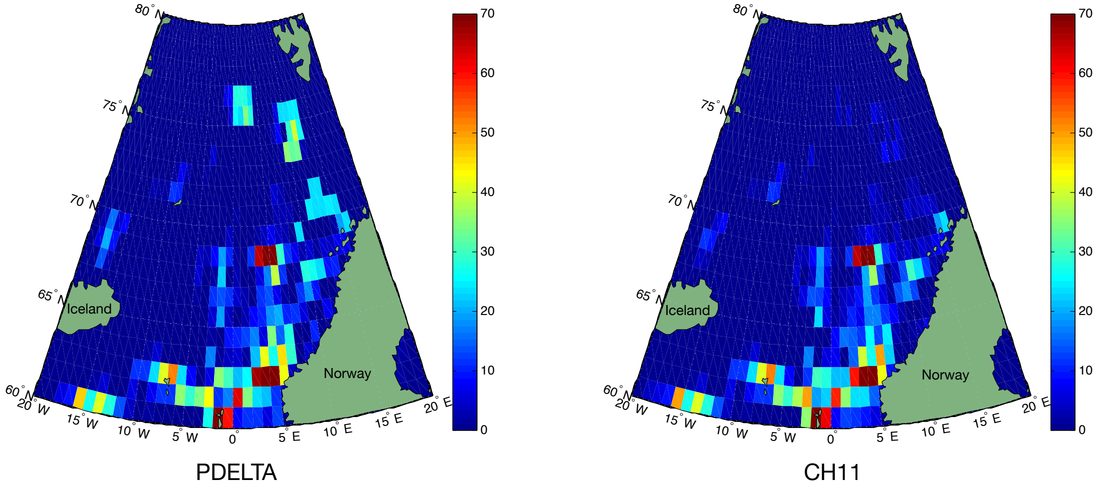


Figure 8.5: Eddy geographic distribution density. The eddies PDELTA identified (left) were similarly distributed as CH11 (right), except for higher latitudes where PDELTA successfully identified a region known for spawning high latitude eddies [205] (see Figure 8.6). Landmasses are colored in dark green.

ignored by CH11's algorithm once the data were filtered. Figure 8.4 shows the monthly cyclonic (top) and anti-cyclonic (bottom) counts for PDELTA (blue curve) and CH11 (red curve). We find that although the counts match well, PDELTA detected fewer eddies than CH11 during winter months, but more eddies during summer months. This may be due to the gaps in satellite data during winter or simply due to seasonality (*i.e.* fewer eddies in winter time).

To visualize PDELTA's performance compared to CH11, we computed the eddy distribution density based on the locations where eddies were detected. We first divided the Nordic Sea region into 1° cells and counted the total number of eddies passing through each cell. The spatial distribution densities are shown in Figure 8.5. Both algorithms detect eddies in "active regions" such as near the Norwegian coast. PDELTA, however, finds more eddies than CH11 in higher latitudes possibly because without data filtering we do not wipeout small-scale eddies. The mean radius of eddies decrease approximately monotonically from about 200 km near the equator to about 75 km at 60° latitude [46, 105], giving our algorithm an advantage to detect such eddies.

As shown in Figure 8.6, the northern region captured by PDELTA and not by

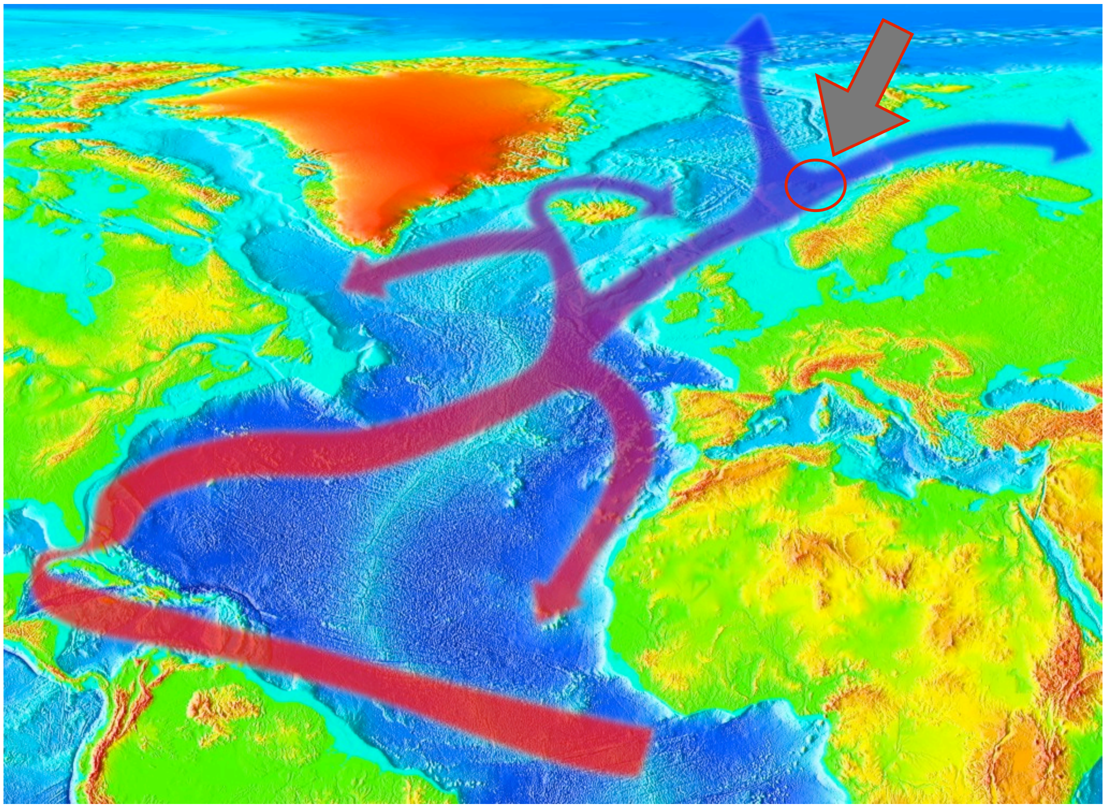


Figure 8.6: Path of the Gulf Stream (red band) from the tropics to the Arctic. The northern eddies that our PDELTA algorithm identifies in Figure 8.5 (big gray arrow) is near the region where the Gulf Stream splits generating high latitude eddies that CH11 failed to capture. The existence of such eddies confirmed by field experiments [205]. Figure source: `http://en.wikipedia.org/wiki/Gulf_Stream`

CH11 (see Figure 8.5) is where the Gulf Stream splits into the North Cape Current (right split) and the West Spitsbergena (left split) [5]. Though different conditions may trigger eddies, they are frequently observed in relation to meandering flows in strong ocean currents (like the Gulf stream [47]). Moreover, the region we identified has been specifically monitored using buoyant floats and was found to have noticeable eddy activity [205]. This is further indication that PDELTA is better able to track small eddies.

Although our algorithm has proven capable of monitoring eddies, the increasing size and resolution of climate data make it imperative that algorithms be scalable for large climate datasets. Therefore, for our algorithm to have real world value, it must scale up to the demands of the climate research community.

To test our algorithm's performance on the potential increase in both data resolution and timespan we compared our algorithm's performance to a connected component eddy searching algorithm similar to the one described in Appendix B of CH11. We constructed two datasets from the original data, one where the data's resolution increased up to 100 times (10 times in each grid dimension) and a second where the length of the time-series increased up to 100 times. We then tracked eddies at weekly intervals for the entire period covered by the data.

Computational Complexity. For a grid with $M \times N$ observations and time-series length K, the time complexity of PDELTA is linear in the number of grid cells, and the time-series length (*i.e.* $O(MNK)$). On the other hand, connected component algorithms are quadratic in the number of grid cells, (*i.e.* $O((MN)^2K)$). The space requirements of both approaches are modest. Figure 8.7 shows empirical results comparing the computation time of PDELTA and the connected component algorithm as the number of grid cells ($M \times N$) and time-series length (K) are increased; the figure shows quadratic increase in computation time for the connected component algorithm as $M \times N$ is increased, while PDELTA's computation time increases linearly. This difference is particularly germane since data from future climate models and satellite observations will be of much higher resolution ($M \times N$ is expected to increase by orders of magnitude) than today's datasets and will approach 100s of petabytes [189].

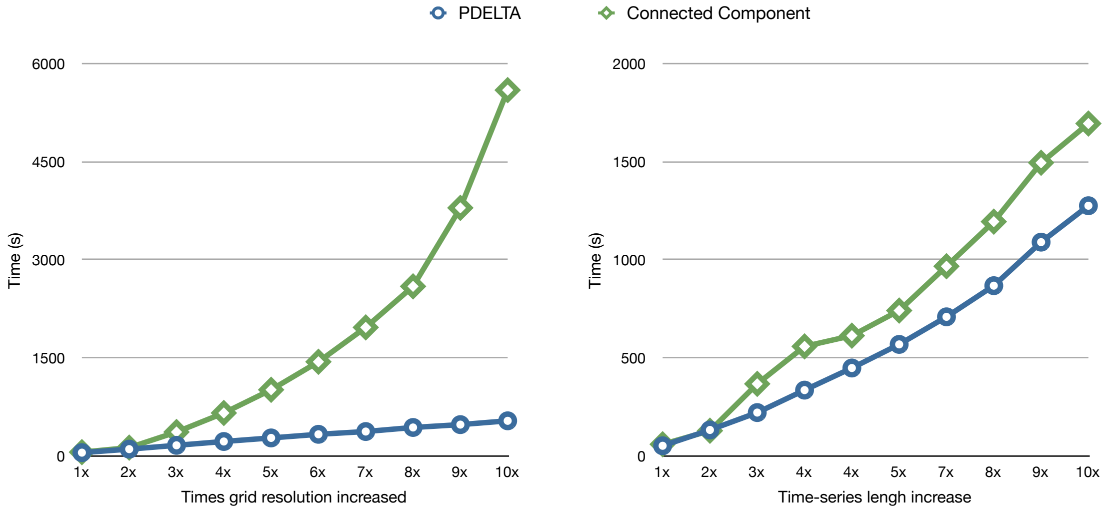


Figure 8.7: Scalability comparison between our algorithm PDELTA (blue) and a connected component algorithm (green) similar to CH11. **Left:** time required to track all eddies in the dataset as a function of the grid resolution. **Right:** time required to track all eddies in the dataset as a function of the time-series length (*i.e.* number of weekly observations). Our algorithm PDELTA (blue) scales better than the connected component algorithm in both time and space.

8.4 Discussion & Future Work

We presented an automated, accurate, and scalable eddy detection and monitoring algorithm based on gradual changes in SSH time-series. While unique, our approach currently suffers from several limitations: First, our minimum and maximum eddy size criteria are hard-coded parameters. Given that we do not filter our data, large-scale phenomena such as gyres and currents are still present in the data and can be mistaken for very large eddies. Conversely, although we are able to detect smaller eddies than CH11, we must ensure that a sufficient number of neighboring time-series experience similar gradual change. To address this issue we imposed a user-specified minimum and maximum eddy size. Future iterations of the algorithm would benefit from automatic parameter estimation that adapts minimum and maximum eddy size based on latitude, since eddy sizes vary as a function of latitude [105]. Another potential investigation could be a sensitivity study of our algorithm's performance to the space and time thresholds mentioned above.

Second, our algorithm does not account for gaps in the data and simply ignores missing values. That is why certain weeks will have very low eddy counts, especially during winter, when precipitation obstructs satellite visibility. Previous works have addressed this challenge by using extrapolation to fill in missing data [48]. Although our algorithm can take advantage of extrapolation, such an approach is not ideal since the factors that impact most climate data such as SSH are highly non-linear, whereas existing extrapolation techniques are not and, therefore, cannot capture such influences. Such a limitation significantly decreases the reliability of data extrapolation. Instead, we would propose to use time-series interpolation algorithms, which have recently been studied in the context of seasonal earth science data [129] as well as established matrix completion techniques [31].

Third, previous studies investigating eddy dynamics have mistaken eddies with Rossby waves (large-scale ocean variability that masquerades as an eddy). Because we analyze global data, and given that we do not filter the data beforehand, there is an increased likelihood of false-positives. To address this issue, CH11 used the eddies' nonlinearity (the ratio of an eddy's rotational and transitional speeds) as a mechanism to differentiate between eddies and Rossby waves. In future work, we will add this type of discriminant to eliminate Rossby waves from our analysis. It is important to note that this issue does not impact our current analysis given that Rossby waves are easily discernible from ocean eddies at high latitudes [105].

Finally, although we report similar monthly eddy counts and spatial distribution as CH11, we must caution that analyzing eddy count alone is an incomplete comparison to CH11. Additional eddy statistics and kinematic properties such as eddy size and speed must be analyzed to fully compare PDELTA to CH11.

Despite these limitations, our algorithm produces similar counts to the state-of-the-art eddy tracking algorithms by leveraging the natural spatio-temporal characteristics of eddies. Additionally, it is capable of identifying regions that other algorithms cannot and these regions have been corroborated by field studies [205]. Identifying small eddies ($<$ 100km) has been challenging for connected component techniques given current data resolution and filtering techniques. By monitoring SSH patterns through changes in the time-series as opposed to visual images, we are able to capture smaller eddies than possible with the existing approaches. Finally, as opposed to common connected component algorithms that run quadratically in the grid resolution, PDELTA runs in linear time – a significant improvement given the expected dramatic increase in climate data [189].

We foresee our algorithm being used in other domains where one is interested in automatically monitoring gradual changes in time-series data. A recent paper by Giles et al. [110] monitored the western Arctic Beaufort Gyre using SSH from satellite data. Automatic gyre monitoring can also be an application to our algorithm. While this line of research is critical to understanding future ocean dynamics and marine ecosystems, it also has important applications to time-series and matrix completion, scalable learning algorithms, and spatio-temporal data mining.

Chapter 9

Conclusion and Future Directions

This thesis introduced the notion of spatio-temporal data mining for climate applications. We presented a broad review of some of the unique characteristics of climate data along with a sample of STDM applications. We encourage interested readers to refer to the references and citations within for further reading. Our survey on TC forecasting models is, as far as we know, the most comprehensive survey on the topic. Our results indicate that effectively incorporating the spatio-temporal context of the data helps improve the performance of tradition data mining tasks.

Based on some of the information presented in this thesis, there may be several traditional data mining concepts that might need rethinking as we explore new applications within spatio-temporal climate data. One such re-thinking might deal with significance testing. The challenge of quantifying statistical significance in climate applications stems from both the exploratory nature of the work as well as a the autocorrelation in the data. While traditional randomization tests (e.g. [170]) may address some of the concerns stemming from multiple hypothesis testing, there is an acute need to develop spatio-temporal randomization test where the randomization procedure does not break the data's inherent characteristics such as autocorrelation. We might also have to re-think the definition of anomalies and extremes beyond that of abnormal deviation from the mean. Climate extremes may be better analyzed in a multi-variate fashion, where multiple relatively normal conditions may lead to a "cumulative" extreme. For instance, while hurricane Katrina was a Category 5 hurricane, it was the breaking of the levee that accentuated its horrific impact. Finally, traditional evaluation metrics for learning algorithms may need to be extended for STDM. A large number of climate problems have no reliable "ground truth" data and therefore rely on unsupervised learning techniques. Hence, it is crucial to develop objective performance measures and experiments that allow to compare the performance of different unsupervised STDM algorithms. Furthermore, traditional performance measures such root mean square error might need to be adjusted to account for spatio-temporal variability.

There are also great opportunities for novel STDM applications within climate science. Within the applications of user-defined pattern mining, the majority of features of interest are usually defined by domain experts. Such an approach is not always feasible since we have significant knowledge gaps in many domains where such data exists. Therefore developing unsupervised feature extraction techniques that autonomously identify significant features based on spatio-temporal variability (*i.e.* how different is a pattern from random noise) might be preferable, especially in large datasets. Additionally, given the large number of climate datasets, each at a different spatio-temporal resolution, there is a high demand for spatio-temporal relationship mining and predictive modeling techniques, that take data at a low, global resolution and infer impact on a higher, local resolution (and vice versa). Finally, one fundamental quantification might need to emerge between uncertainty and risk. Data mining and machine learning have used probabilities as a measure of uncertainty. However, numerous climate-related questions are interested in risk as opposed to uncertainty. Providing decision-makers with tools to convert statistical uncertainty to risk quantities based on available information is has the potential to be a major scientific and societal contribution.

Answers to some of these questions will emerge over time as we continue to see new STDM applications to climate data. Others, such as significance tests, might require diligent collaborations with adjacent fields such as statistics. Nonetheless, there is an exciting (and challenging) road ahead for STDM researchers.

Bibliography

[1] G. a. Meehl and T. F. Stocker. Global climate projections. In S. Solomon, M. Manning, Z. Chen, M. Marquis, K. Averyt, M. Tignor, and H. Miller, editors, *Climate Change 2007: the Physical Science Basis. Contribution of Working Group 1 to the Fourth Assesment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press, Cambridge, Uk and New York, Ny, Usa, 2007.

[2] R. a. Pielke. Are there trends in hurricane destruction? *Nature*, 438(7071):E11, Dec. 2005. ISSN 0028-0836. doi: 10.1038/Nature04426. URL `Http://Dx.Doi.Org/10.1038/Nature04426`.

[3] G. a Vecchi, K. L. Swanson, and B. J. Soden. Whither hurricane activity. *Science*, 322(5902):687, 2008.

[4] S. D. Aberson. Regimes or cycles in tropical cyclone activity in the north atlantic. *Bulletin of the American Meteorological Society*, pages 39–43, 2009.

[5] A. AMAP. Assessment report: Arctic pollution issues. *Arctic Monitoring and Assessment Programme (AMAP), Oslo, Norway*, 12:859, 1998.

[6] T. C. B. Anbaroğlu. Spatio-temporal outlier detection in environmental data. *Spatial and Temporal Reasoning for Ambient Intelligence Systems*, pages 1–9, 2009.

[7] A. Andrew. Another efficient algorithm for convex hulls in two dimensions. *Information Processing Letters*, 9(5):216–219, 1979.

[8] A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno, and C. Zhou. Synchronization in complex networks. *Physics Reports*, 469(3):93–153, 2008.

[9] K. Ashok and T. Yamagata. The el niño with a difference. *Nature*, 461(7263), 2009.

[10] K. Ashok, S. Behera, S. Rao, H. Weng, and T. Yamagata. El niño modoki and its possible teleconnection. *J. Geophys. Res*, 112(10.1029), 2007.

[11] C. L. Bain, J. De Paz, J. Kramer, G. Magnusdottir, P. Smyth, H. Stern, and C.-c. Wang. Detecting the itcz in instantaneous satellite data using spatiotemporal statistical modeling: Itcz climatology in the east pacific. *Journal of Climate*, 24 (1):216–230, 2011.

[12] D. Baldocchi, E. Falge, L. Gu, R. Olson, D. Hollinger, S. Running, P. Anthoni, C. Bernhofer, K. Davis, R. Evans, et al. Fluxnet: a new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological Society*, 82(11):2415–2434, 2001.

[13] Y. Bar-Shalom and X. Li. *Multitarget-multisensor tracking: principles and techniques*. YBS Publishing, Storrs, CT: University of Connecticut, 1995.

[14] Y. Bar-Shalom and E. Tse. Tracking in a cluttered environment with probabilistic data association. *Automatica*, 11(5):451–460, 1975.

[15] S. Barua and R. Alhajj. Parallel wavelet transform for spatio-temporal outlier detection in large meteorological data. *Intelligent Data Engineering and Automated Learning-IDEAL 2007*, pages 684–694, 2007.

[16] J. Basak, A. Sudarshan, D. Trivedi, and M. Santhanam. Weather data mining using independent component analysis. *The Journal of Machine Learning Research*, 5:239–253, 2004.

[17] J. Belanger, J. Curry, and P. Webster. Predictability of north atlantic tropical cyclone activity on intraseasonal time scales. *Monthly Weather Review*, 138(12): 4362–4374, 2010.

[18] J. Belanger, P. Webster, and J. Curry. Extended prediction of north indian ocean tropical cyclones using the extended prediction of north indian ocean tropical

cyclones using the ecmwf variable ensemble prediction system. *Mon. Wea. Rev*, 2011.

[19] M. a. Bender, T. R. Knutson, R. E. Tuleya, J. J. Sirutis, G. a. Vecchi, S. T. Garner, and I. M. Held. Modeled Impact of Anthropogenic Warming on the Frequency of Intense Atlantic Hurricanes. *Science*, 327(5964):454–458, 2010. doi: 10.1126/Science.1180568. URL `Http://Www.Sciencemag.Org/Content/327/5964/454.Abstract`.

[20] L. Bengtsson, K. I. Hodges, M. Esch, N. Keenlyside, L. Kornblueh, J. J. Luo, and T. Yamagata. How may tropical cyclones change in a warmer climate? *Tellus a*, 59(4):539561, 2007. ISSN 1600-0870.

[21] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, pages 1165–1188, 2001.

[22] Y. Berezin, A. Gozolchiani, O. Guez, and S. Havlin. Stability of climate networks with time. *Scientific Reports*, 2, 2012.

[23] M. Bister and K. Emanuel. Dissipative heating and hurricane intensity. *Meteorology and Atmospheric Physics*, 65(3):233–240, 1998.

[24] S. Blackman. Multiple hypothesis tracking for multiple target tracking. *Aerospace and Electronic Systems Magazine, IEEE*, 19(1):5–18, 2004.

[25] E. Blake and W. Gray. Prediction of august atlantic basin hurricane activity. *Weather and Forecasting*, 19(6):1044–1060, 2004.

[26] S. Boriah, V. Kumar, M. Steinbach, C. Potter, and S. Klooster. Land cover change detection: a case study. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 857–865. ACM, 2008.

[27] A. Braverman and E. Fetzer. Mining massive earth science data sets for large scale structure. In *Proceedings of the Earth-Sun System Technology Conference*, 2005.

[28] S. Camargo, a. Barnston, P. Klotzbach, and C. Landsea. Seasonal tropical cyclone forecasts. *Wmo Bulletin*, 56(4):297, 2007.

[29] S. J. Camargo, A. W. Robertson, S. J. Gaffney, P. Smyth, and M. Ghil. Cluster analysis of typhoon tracks. part i: General properties. *Journal of Climate*, 20(14): 3635–3653, 2007.

[30] S. J. Camargo, A. W. Robertson, S. J. Gaffney, P. Smyth, and M. Ghil. Cluster analysis of typhoon tracks. part ii: Large-scale circulation and enso. *Journal of climate*, 20(14):3654–3676, 2007.

[31] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[32] X. Capet, P. Klein, B. Hua, G. Lapeyre, and J. McWilliams. Surface kinetic energy transfer in surface quasi-geostrophic flows. *Journal of Fluid Mechanics*, 604(1):165–174, 2008.

[33] T. Carlson. Weather note: an apparent relationship between the sea-surface temperature of the tropical atlantic and the development of african disturbances into tropical storms. *Monthly Weather Review*, 99(4):309–310, 1971.

[34] M. Castellani. Identification of eddies from sea surface temperature maps with neural networks. *International journal of remote sensing*, 27(8):1601–1618, 2006.

[35] A. Chaigneau and O. Pizarro. Mean surface circulation and mesoscale turbulent flow characteristics in the eastern south pacific from satellite tracked drifters. *J. Geophys. Res*, 110:C05014, 2005.

[36] A. Chaigneau, A. Gizolme, and C. Grados. Mesoscale eddies off peru in altimeter records: Identification algorithms and eddy spatio-temporal patterns. *Progress in Oceanography*, 79(2-4):106–119, 2008.

[37] T. Cham and J. Rehg. A multiple hypothesis approach to figure tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2. IEEE, 1999.

[38] Y. Chamber, A. Garg, V. Mithal, I. Brugere, M. Lau, V. Krishna, S. Boriah, M. Steinbach, V. Kumar, C. Potter, and S. A. Klooster. A novel time series based approach to detect gradual vegetation changes in forests. In *CIDU 2011:*

Proceedings of the NASA Conference on Intelligent Data Understanding, pages 248–262, 2011.

[39] J. Chan. Comment on" changes in tropical cyclone number, duration, and intensity in a warming environment". *Science*, 311(5768):1713, 2006.

[40] J. Chan, J. Shi, and C. Lam. Seasonal forecasting of tropical cyclone activity over the western north pacific and the south china sea. *Weather Forecast*, 13(4): 997–1004, 1998.

[41] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection for discrete sequences: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 24 (5):823–839, 2012.

[42] E. K. M. Chang and Y. Guo. Is the number of north atlantic tropical cyclones significantly underestimated prior to the availability of satellite observations? *Geophysical Research Letters*, 34:5 Pp., July 2007. doi: 200710.1029/2007Gl030169. URL `Http://Www.Agu.Org.Floyd.Lib.Umn.Edu/Pubs/Crossref/2007/2007Gl030169.Shtml`.

[43] S. Chatterjee, K. Steinhaeuser, A. Banerjee, S. Chatterjee, and A. Ganguly. Sparse group lasso: Consistency and climate applications. SDM, 2012.

[44] F. Chauvin, J. Royer, and M. Deque. Response of hurricane-type vortices to global warming as simulated by arpege-climat at high resolution. *Climate Dynamics*, 27 (4):377–399, 2006.

[45] M. Chelliah and G. D. Bell. Tropical multidecadal and interannual climate variability in the ncep–ncar reanalysis. *Journal of Climate*, 17(9):1777–1803, 2011/07/27 2004. doi: 10.1175/1520-0442(2004)017⟨1777:Tmaicv⟩2.0.Co;2. URL `Http://Dx.Doi.Org/10.1175/1520-0442(2004)017<1777:Tmaicv>2.0.Co;2`.

[46] D. Chelton, M. Schlax, R. Samelson, and R. de Szoeke. Global observations of large oceanic eddies. *Geophysical Research Letters*, 34:L15606, 2007.

[47] D. Chelton, M. Schlax, and R. Samelson. Global observations of nonlinear mesoscale eddies. *Progress in Oceanography*, 2011.

[48] D. B. Chelton, P. Gaube, M. G. Schlax, J. J. Early, and R. M. Samelson. The influence of nonlinear mesoscale eddies on near-surface oceanic chlorophyll. *Science*, 334(6054):328–332, 2011.

[49] S. Chen, W. Zhao, M. Donelan, J. Price, and E. Walsh. the cblast-hurricane program and the next-generation fully coupled atmosphere–wave–ocean models for hurricane research and prediction. 2007.

[50] Y. Chen, J. T. Randerson, D. C. Morton, R. S. DeFries, G. J. Collatz, P. S. Kasibhatla, L. Giglio, Y. Jin, and M. E. Marlier. Forecasting fire season severity in south america using sea surface temperature anomalies. *Science*, 334(6057): 787–791, 2011.

[51] T. Cheng and Z. Li. A multiscale approach for spatio-temporal outlier detection. *Transactions in GIS*, 10(2):253–263, 2006.

[52] P. A. Chou, T. Lookabaugh, and R. M. Gray. Entropy-constrained vector quantization. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 37(1): 31–42, 1989.

[53] P. Chu and X. Zhao. A bayesian regression approach for predicting seasonal tropical cyclone activity over the central north pacific. *Journal of Climate*, 20 (15):4002–4013, 2007.

[54] P. Chu, X. Zhao, C. Lee, and M. Lu. Climate prediction of tropical cyclone activity in the vicinity of taiwan using the multivariate least absolute deviation regression method. *Terrestrial Atmospheric and Oceanic Sciences*, 18(4):805, 2007.

[55] J. Cl Chan, J. Shi, and K. Liu. Improvements in the seasonal forecasting of tropical cyclone activity over the western north pacific. *Weather and Forecasting*, 16(4):491–498, 2001.

[56] P. Clark and S. Matwin. Using qualitative models to guide inductive learning. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 49–56, 1993.

[57] S. H. Clearwater and F. J. Provost. Rl4: A tool for knowledge-based induction. In *Tools for Artificial Intelligence, 1990., Proceedings of the 2nd International IEEE Conference on*, pages 24–30. IEEE, 1990.

[58] R. Coe and R. Stern. Fitting models to daily rainfall data. *Journal of Applied Meteorology*, 21(7):1024–1031, 1982.

[59] M. Collins, T. Fricker, and L. Hermanson. From observations to forecasts–part 9: What is decadal forecasting? *Weather*, 66(6):160–164, 2011.

[60] D. Cox and V. Isham. A simple spatial-temporal model of rainfall. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 415(1849): 317–328, 1988.

[61] I. Cox and S. Hingorani. An efficient implementation of reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(2):138–150, 1996.

[62] N. Cressie and C. K. Wikle. *Statistics for spatio-temporal data*, volume 465. Wiley, 2011.

[63] N. Cressie, R. Assunçao, S. H. Holan, M. Levine, J. Zhang, and C.-N. SAMSI. Dynamical random-set modeling of concentrated precipitation in north america. *Statistics and its Interface*, 2011.

[64] D. D'Alimonte. Detection of mesoscale eddy-related structures through iso-sst patterns. *Geoscience and Remote Sensing Letters, IEEE*, 6(2):189–193, 2009.

[65] M. Demaria and J. Kaplan. an updated statistical hurricane intensity prediction scheme (ships) for the atlantic and eastern north pacific basins. *Weather and Forecasting*, 14(3):326–337, 1999.

[66] M. Demaria, J. Knaff, and B. Connell. A tropical cyclone genesis parameter for the tropical atlantic. *Weather and Forecasting*, 16(2):219–233, 2001.

[67] M. Demaria, M. Mainelli, L. Shay, J. Knaff, and J. Kaplan. Further improvements to the statistical hurricane intensity prediction scheme (ships). *Weather and Forecasting*, 20(4):531–543, 2005.

[68] M. Diaz. Isentropic descent beneath the saharan air layer and its impact on tropical cyclogenesis, 2009.

[69] M. Diaz and F. Semazzi. *the Role of West African Coastal Upwelling in the Genesis of Tropical Cyclones: a New Mechanism.* 2008. Published: Newsletter of the Climate Variability and Predictability Programme.

[70] Y. H. Ding and E. R. Reiter. *Large-Scale Circulation Conditions Affecting the Variability in the Frequency of Tropical Cyclone Formation over the North Atlantic and the North Pacific Oceans*. Colorado State University, 1981.

[71] P. Domingos. Occam's two razors: The sharp and the blunt. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 37–43. AAAI Press, 1998.

[72] C. Dong, F. Nencioli, Y. Liu, and J. McWilliams. An automated approach to detect oceanic eddies from satellite remotely sensed sea surface temperature data. *Geoscience and Remote Sensing Letters, IEEE*, (99):1–5, 2011.

[73] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. The backbone of the climate network. *EPL (Europhysics Letters)*, 87(4):48007, 2009.

[74] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. Complex networks in climate dynamics. *The European Physical Journal-Special Topics*, 174(1):157–179, 2009.

[75] B. Effron and R. Tibshirani. Statistical data analysis in the computer age. *Science*, 253(5018):390–395, 1991.

[76] J. Elsner and T. Jagger. a Hierarchical Bayesian Approach to Seasonal Hurricane Modeling. *J. Climate*, 17:2813–2827, 2004.

[77] J. Elsner and T. Jagger. Prediction models for annual u.s. hurricane counts. *Journal of Climate*, 19(12):2935–2952, 2006. doi: 10.1175/Jcli3729.1. URL `Http://Journals.Ametsoc.Org/Doi/Abs/10.1175/Jcli3729.1`.

[78] J. Elsner and C. Schmertmann. Improving extended-range seasonal predictions of intense atlantic hurricane activity. *Weather and Forecasting*, 8(3):345–351, 1993.

[79] J. Elsner, X. Niu, and A. Tsonis. Multi-year prediction model of north atlantic hurricane activity. *Meteorology and Atmospheric Physics*, 68(1):43–51, 1998.

[80] J. Elsner, T. Jagger, M. Dickinson, and D. Rowe. Improving multiseason forecasts of north atlantic hurricane activity. *Journal of Climate*, 21(6):1209–1219, 2008. doi: 10.1175/2007Jcli1731.1. URL `Http://Journals.Ametsoc.Org/Doi/Abs/10.1175/2007Jcli1731.1`.

[81] J. Elsner, J. Kossin, and T. Jagger. the increasing intensity of the strongest tropical cyclones. *Nature*, 455(7209):92–95, 2008.

[82] J. Elsner, T. Jagger, and E. Fogarty. Visibility network of united states hurricanes. *Geophysical Research Letters*, 36(16):L16702, 2009.

[83] J. Elsner, T. Jagger, and R. Hodges. Daily tropical cyclone intensity response to solar ultraviolet radiation. *Geophys. Res. Lett*, 37:L09701, 2010.

[84] K. Emanuel. the dependence of hurricane intensity on climate. *Nature*, 326(6112):483–485, 1987.

[85] K. Emanuel. Contribution of tropical cyclones to meridional heat transport by the oceans. *Journal of Geophysical Research*, 106:14, 2001.

[86] K. Emanuel. Tropical cyclones. *Annual Review of Earth and Planetary Sciences*, 31(1):75, 2003.

[87] K. Emanuel. Increasing destructiveness of tropical cyclones over the past 30 years. *Nature*, 436(7051):686–688, 2005. ISSN 0028-0836. doi: 10.1038/Nature03906. URL `Http://Dx.Doi.Org.Floyd.Lib.Umn.Edu/10.1038/Nature03906`.

[88] K. Emanuel. the hurricane-climate connection. *Bulletin of the American Meteorological Society*, 89(5), 2008. ISSN 1520-0477.

[89] K. Emanuel and D. Nolan. Tropical cyclone activity and the global climate system. In *Proc. 26Th Conf. on Hurricanes and Tropical Meteorology*, 2004.

[90] K. Emanuel, R. Sundararajan, and J. Williams. Hurricanes and Global Warming. *Bull. Am. Meteorol. Soc*, 89:347–367, 2008.

[91] R. Eppley and B. Peterson. Particulate organic matter flux and planktonic new production in the deep ocean. *Nature*, 282(5740):677–680, 1979.

[92] E. Epstein. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8:985–987, 1969.

[93] a. Evan, J. Dunion, J. Foley, a. Heidinger, and C. Velden. New evidence for a relationship between atlantic tropical cyclone activity and african dust outbreaks. *Geophys. Res. Lett*, 33:L19813, 2006.

[94] J. Faghmous, Y. Chamber, F. Vikebø, S. Boriah, S. Liess, M. d.S. Mesquita, and V. Kumar. A novel and scalable spatio-temporal technique for ocean eddy monitoring. In *Twenty-Sixth Conference on Artificial Intelligence (AAAI-12)*, 2012.

[95] J. H. Faghmous, L. Styles, V. Mithal, S. Boriah, S. Liess, F. Vikebo, M. d. S. Mesquita, and V. Kumar. Eddyscan: A physically consistent ocean eddy monitoring application. In *Intelligent Data Understanding (CIDU), 2012 Conference on*, pages 96 –103, oct. 2012.

[96] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. *Fast subsequence matching in time-series databases*, volume 23. ACM, 1994.

[97] K. Fan. A prediction model for atlantic named storm frequency using a year-by-year increment approach. *Weather and Forecasting*, 25(6):1842–1851, 2010.

[98] F. Fang and R. Morrow. Evolution, movement and decay of warm-core leeuwin current eddies. *Deep Sea Research Part II: Topical Studies in Oceanography*, 50 (12-13):2245–2261, 2003.

[99] A. Fernandes. Identification of oceanic eddies in satellite images. *Advances in Visual Computing*, pages 65–74, 2008.

[100] A. Fernandes and S. Nascimento. Automatic water eddy detection in sst maps using random ellipse fitting and vectorial fields for image segmentation. In *Discovery Science*, pages 77–88. Springer, 2006.

[101] E. Fisher. Hurricanes and the sea-surface temperature field. *Journal of Atmospheric Sciences*, 15:328–333, 1958.

[102] E. A. Fogarty, J. B. Elsner, T. H. Jagger, and A. A. Tsonis. Network analysis of us hurricanes. *Hurricanes and Climate Change*, pages 153–167, 2009.

[103] J. A. Foley. Can we feed the world & sustain the planet? *Scientific American*, 305(5):60–65, 2011.

[104] N. Frank. Atlantic tropical systems of 1974. *Monthly Weather Review*, 103:294, 1975.

[105] L. Fu, D. Chelton, P. Le Traon, and R. Morrow. Eddy dynamics from satellite altimetry. *Oceanography*, 23(4):14–25, 2010.

[106] Q. Fu, A. Banerjee, S. Liess, and P. K. Snyder. Drought detection of the last century: An mrf-based approach. In *Proceedings of the SIAM International Conference on Data Mining*, 2012.

[107] S. J. Gaffney, A. W. Robertson, P. Smyth, S. J. Camargo, and M. Ghil. Probabilistic clustering of extratropical cyclones using regression mixture models. *Climate Dynamics*, 29(4):423–440, 2007.

[108] S. Garner, I. Held, T. Knutson, and J. Sirutis. the roles of wind shear and thermal stratification in past and projected changes of atlantic tropical cyclone activity. *Journal of Climate*, 22(17):4723–4734, 2009.

[109] S. Ghosh, D. Das, S.-C. Kao, and A. R. Ganguly. Lack of uniform trends but increasing spatial variability in observed indian rainfall extremes. *Nature Climate Change*, 2011.

[110] K. A. Giles, S. W. Laxon, A. L. Ridout, D. J. Wingham, and S. Bacon. Western arctic ocean freshwater storage increased by wind-driven spin-up of the beaufort gyre. *Nature Geosci*, 2012.

[111] S. Goldenberg and L. Shapiro. Physical mechanisms for the association of el niño and west african rainfall with atlantic major hurricane activity. *Journal of Climate*, 9(6):1169–1187, 1996.

[112] S. Goldenberg, C. Landsea, A. Mestas-Nuñez, and W. Gray. The recent increase in atlantic hurricane activity: Causes and implications. *Science*, 293(5529):474, 2001.

[113] D. L. Gonzalez, Z. Chen, T. Pansombut, F. Semazzi, V. Kumar, a. Melechko, and N. Samatova. Hierarchical classifier-regressor ensemble for multi-phase non-linear dynamic system response prediction: Application to hurricane activity estimation. Technical report, North Carolina State University, 2011.

[114] W. Gray. Hurricanes: Their formation, structure and likely role in the tropical circulation. *Meteorology over the Tropical Oceans*, pages 155–218, 1979.

[115] W. Gray. Atlantic seasonal hurricane frequency. part i: El niño and 30 mb quasi-biennial oscillation influences. *Mon. Wea. Rev*, 112(9):1649–1668, 1984.

[116] W. Gray. Atlantic seasonal hurricane frequency. part ii: Forecasting its variability. *Mon. Wea. Rev*, 112(9):1669–1683, 1984.

[117] W. Gray. Environmental influences on tropical cyclones. *Australian Meteorological Magazine*, 36(3):127–39, 1988.

[118] W. Gray, C. Landsea, P. Mielke, and K. Berry. Predicting atlantic seasonal hurricane activity 6-11 months in advance. *Weather and Forecasting*, 7(3):440–455, 1992.

[119] W. M. Gray. Global view of the origin of tropical disturbances and storms. *Monthly Weather Review*, 96(10):669700, 1968. ISSN 1520-0493.

[120] S. Gualdi, E. Scoccimarro, a. Navarra, and Others. Changes in tropical cyclone activity due to global warming: Results from a high-resolution coupled general circulation model. *Journal of Climate*, 21(20):5204–5228, 2008.

[121] O. Guez, A. Gozolchiani, Y. Berezin, S. Brenner, and S. Havlin. Climate network structure evolves with north atlantic oscillation phases. *EPL (Europhysics Letters)*, 98(3):38006, 2012.

[122] M. Han, W. Xu, H. Tao, and Y. Gong. An algorithm for multiple object trajectory tracking. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004.*

Proceedings of the 2004 IEEE Computer Society Conference on, volume 1, pages I–864. IEEE, 2004.

[123] C. Hansen, E. Kvaleberg, and A. Samuelsen. Anticyclonic eddies in the norwegian sea; their generation, evolution and impact on primary production. *Deep Sea Research Part I: Oceanographic Research Papers*, 57(9):1079–1091, Sept. 2010. ISSN 0967-0637. doi: 10.1016/j.dsr.2010.05.013. URL `http://www.sciencedirect.com/science/article/B6VGB-5086090-1/2/b2ea6387d23b35587496c7fb86f654bb`.

[124] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[125] I. M. Held and B. J. Soden. Robust responses of the hydrological cycle to global warming. *Journal of Climate*, 19(21):56865699, 2006. ISSN 1520-0442.

[126] D. Henke, P. Smyth, C. Haffke, and G. Magnusdottir. Automated analysis of the temporal behavior of the double intertropical convergence zone over the east pacific. *Remote Sensing of Environment*, 123:418–433, 2012.

[127] C. Hennon and J. Hobgood. Forecasting Tropical Cyclogenesis over the Atlantic Basin Using Large-Scale Data. *Monthly Weather Review*, 131(12):2927–2940, 2003. ISSN 1520-0493.

[128] C. Hennon, C. Marzban, and J. Hobgood. Improving tropical cyclogenesis statistical model forecasts through the application of a neural network classifier. *Weather and Forecasting*, 20(6):1073–1083, 2005.

[129] J. Hird and G. McDermid. Noise reduction of ndvi time-series: An empirical comparison of selected techniques. *Remote Sensing of Environment*, 113(1):248 – 258, 2009.

[130] F. M. Hoffman, W. W. Hargrove Jr, D. J. Erickson III, and R. J. Oglesby. Using clustered climate regimes to analyze and compare predictions from fully coupled general circulation models. *Earth Interactions*, 9(10):1–27, 2005.

[131] G. Holland and P. Webster. Heightened Tropical Cyclone Activity in the North Atlantic: Natural Variability or Climate Trend? *Philosophical Transactions of the Royal Society a: Mathematical, Physical and Engineering Sciences*, 365(1860): 2695, 2007. ISSN 1364-503X.

[132] G. J. Holland. the maximum potential intensity of tropical cyclones. *Journal of the Atmospheric Sciences*, 54(21):25192541, 1997. ISSN 1520-0469.

[133] R. Holyer and S. Peckinpaugh. Edge detection applied to satellite imagery of the oceans. *Geoscience and Remote Sensing, IEEE Transactions on*, 27(1):46–56, 1989.

[134] S. Hopsch, C. Thorncroft, K. Hodges, and a. Aiyyer. West African Storm Tracks and Their Relationship to Atlantic Tropical Cyclones. *Journal of Climate*, 20(11): 2468–2483, 2007. ISSN 1520-0442.

[135] S. Hopsch, C. Thorncroft, and K. Tyle. Analysis of african easterly wave structures and their role in influencing tropical cyclogenesis. *Monthly Weather Review*, 2009.

[136] C. Hoyos, P. Agudelo, P. Webster, and J. Curry. Deconvolution of the factors contributing to the increase in global hurricane intensity. *Science*, 312(5770):94, 2006.

[137] H.-C. Huang and N. Cressie. Spatio-temporal prediction of snow water equivalent using the kalman filter. *Computational Statistics & Data Analysis*, 22(2):159–175, 1996.

[138] J. Isern-Fontanet, E. García-Ladona, and J. Font. Identification of marine eddies from altimetric maps. *Journal of Atmospheric and Oceanic Technology*, 20(5): 772–778, 2003.

[139] J. Isern-Fontanet, J. Font, E. García-Ladona, M. Emelianov, C. Millot, and I. Taupier-Letage. Spatial structure of anticyclonic eddies in the algerian basin (mediterranean sea) analyzed using the okubo–weiss parameter. *Deep Sea Research Part II: Topical Studies in Oceanography*, 51(25):3009–3028, 2004.

[140] J. Isern-Fontanet, E. García-Ladona, and J. Font. Vortices of the mediterranean sea: An altimetric perspective. *Journal of physical oceanography*, 36(1):87–103, 2006.

[141] J. Isern-Fontanet, E. García-Ladona, J. Font, and A. García-Olivares. Non-gaussian velocity probability density functions: An altimetric perspective of the mediterranean sea. *Journal of physical oceanography*, 36(11):2153–2164, 2006.

[142] H. Kao and J. Yu. Contrasting Eastern-Pacific and Central-Pacific Types of ENSO. *Journal of Climate*, 22(3):615–632, 2009.

[143] A. Karpatne, M. Blank, M. Lau, S. Boriah, K. Steinhaeuser, M. Steinbach, and V. Kumar. Importance of vegetation type in forest cover estimation. In *CIDU*, pages 71–78, 2012.

[144] J. Kawale, M. Steinbach, and V. Kumar. Discovering dynamic dipoles in climate data. In *SIAM International Conference on Data mining, SDM. SIAM*, 2011.

[145] J. Kawale, S. Chatterjee, D. Ormsby, K. Steinhaeuser, S. Liess, and V. Kumar. Testing the significance of spatio-temporal teleconnection patterns. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 642–650. ACM, 2012.

[146] M. Kim and J. Han. A particle-and-density based evolutionary clustering method for dynamic networks. *Proceedings of the VLDB Endowment*, 2(1):622–633, 2009.

[147] P. Klotzbach. Trends in global tropical cyclone activity over the past twenty years. *Geophys. Res. Lett*, 33:L10805, 2006.

[148] P. Klotzbach. The influence of el niÑo–southern oscillation and the atlantic multi-decadal oscillation on caribbean tropical cyclone activity. *J. Climate*, 24:721–731, 2011.

[149] P. Klotzbach and W. Gray. Twenty-five years of atlantic basin seasonal hurricane forecasts. *Geophysical Research Letters*, 36:L09711, 2009.

[150] P. J. Klotzbach and W. M. Gray. Forecasting september atlantic basin tropical cyclone activity. *Weather and Forecasting*, 18(6):1109–1128, 2011/07/23 2003. URL `Http://Dx.Doi.Org/10.1175/1520-0434(2003)018<1109:Fsabtc>2.0.Co;2`.

[151] P. J. Klotzbach and W. M. Gray. Updated 6–11-month prediction of atlantic basin seasonal hurricane activity. *Weather and Forecasting*, 19(5):917–934, 2011/07/23 2004. URL `Http://Dx.Doi.Org/10.1175/1520-0434(2004)019<0917:Umpoab>2.0.Co;2`.

[152] T. Knutson and R. Tuleya. Impact of co2-induced warming on simulated hurricane intensity and precipitation: Sensitivity to the choice of climate model and convective parameterization. *Journal of Climate*, 17(18):3477–3495, 2004.

[153] T. Knutson, J. Sirutis, S. Garner, I. Held, and R. Tuleya. Simulation of the recent multidecadal increase of atlantic hurricane activity using an 18-km-grid regional model. *Bulletin of the American Meteorological Society*, 88(10):1549–1565, 2007.

[154] T. Knutson, J. Sirutis, S. Garner, G. Vecchi, and I. Held. Simulated Reduction in Atlantic Hurricane Frequency Under Twenty-First-Century Warming Conditions. *Nature Geoscience*, 1(6):359–364, 2008. ISSN 1752-0894.

[155] T. R. Knutson, J. L. Mcbride, J. Chan, K. Emanuel, G. Holland, C. Landsea, I. Held, J. P. Kossin, a. K. Srivastava, and M. Sugi. Tropical cyclones and climate change. *Nature Geoscience*, 2010. ISSN 1752-0894.

[156] A. Kolmogorov. The local structure of turbulence in incompressible viscous fluid for very large reynolds numbers. *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences*, 434(1890):9–13, 1991.

[157] A. N. Kolmogorov. On degeneration of isotropic turbulence in an incompressible viscous liquid. In *Dokl. Akad. Nauk SSSR*, volume 31, pages 538–540, 1941.

[158] J. Kossin, S. Camargo, and M. Sitkowski. Climate Modulation of North Atlantic Hurricane Tracks. *Journal of Climate*, 23(11):3057–3076, June 2010. ISSN 1520-0442.

[159] J. Kug, F. Jin, and S. an. Two Types of El Niño events: Cold Tongue El Niño and Warm Pool El Niño. *Journal of Climate*, 22(6):1499–1515, 2009.

[160] T. Kurien. Issues in the design of practical multitarget tracking algorithms. *Multitarget-Multisensor Tracking: Advanced Applications*, 1:43–83, 1990.

[161] P. J. Lamb and R. A. Peppler. North atlantic oscillation: Concept and an application. *Bulletin of the American Meteorological Society*, 68:1218–1225, 1987.

[162] C. Landsea. Meteorology: Hurricanes and global warming. *Nature*, 438(7071): E11–E12, 2005.

[163] C. Landsea, N. Aeronautics, and D. Space Administration, Washington. a climatology of intense (or major) atlantic hurricanes. *Monthly Weather Review*, 121(6): 1703–1713, 1993.

[164] C. Landsea, G. Bell, W. Gray, and S. Goldenberg. The extremely active 1995 atlantic hurricane season: Environmental conditions and verification of seasonal forecasts. *Monthly Weather Review*, 126(5):1174–1193, 1998.

[165] C. W. Landsea. Counting atlantic tropical cyclones back to 1900. *Eos Trans. Agu*, 88(18):197–202, 2007.

[166] T. Larow, Y. Lim, D. Shin, E. Chassignet, and S. Cocke. Atlantic basin seasonal hurricane simulations. *Journal of Climate*, 21(13):3191–3206, 2008.

[167] S. Laxman and P. S. Sastry. A survey of temporal data mining. *Sadhana*, 31(2): 173–198, 2006.

[168] Y. Lee, B. G. Buchanan, and J. M. Aronis. Knowledge-based learning in exploratory science: Learning rules to predict rodent carcinogenicity. *Machine Learning*, 30(2):217–240, 1998.

[169] I. Lin, W. Liu, C. Wu, G. Wong, C. Hu, Z. Chen, W. Liang, Y. Yang, and K. Liu. New evidence for enhanced ocean primary production triggered by tropical cyclone. *Geophys. Res. Lett*, 30(13):1718, 2003.

[170] R. Livezey and W. Chen. Statistical field significance and its determination by monte carlo techniques(in meteorology). *Monthly Weather Review*, 111:46–59, 1983.

[171] M. Mann and K. Emanuel. Atlantic Hurricane Trends Linked to Climate Change. *Eos*, 87(24):233–244, 2006.

[172] M. Mann, T. Sabbatelli, and U. Neu. Evidence for a modest undercount bias in early historical atlantic tropical cyclone counts. *Geophys. Res. Lett*, 34:L22707, 2007.

[173] D. McGillicuddy Jr. Eddies masquerade as planetary waves. *Science*, 334(6054): 318–319, 2011.

[174] M. McGuire, V. Janeja, and A. Gangopadhyay. Spatiotemporal neighborhood discovery for sensor data. *Knowledge Discovery from Sensor Data*, pages 203–225, 2010.

[175] E. Mesrobian, R. Muntz, E. Shek, J. Santos, J. Yi, K. Ng, S.-Y. Chien, C. Mechoso, J. Farrara, P. Stolorz, et al. Exploratory data mining and analysis using conquest. In *Communications, Computers, and Signal Processing, 1995. Proceedings., IEEE Pacific Rim Conference on*, pages 281–286. IEEE, 1995.

[176] E. Mesrobian, R. Muntz, E. Shek, S. Nittel, M. La Rouche, M. Kriguer, C. Mechoso, J. Farrara, P. Stolorz, and H. Nakamura. Mining geophysical data for knowledge. *IEEE Expert*, 11(5):34–44, 1996.

[177] A. M. Mestas-Nuñez and D. B. Enfield. Rotated global modes of non-enso sea surface temperature variability. *Journal of Climate*, 12(9):2734–2746, 1999.

[178] V. Mithal, A. Garg, S. Boriah, M. Steinbach, V. Kumar, C. Potter, S. Klooster, and J. C. Castilla-Rubio. Monitoring global forest cover using data mining. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(4):36, 2011.

[179] V. Mithal, A. Garg, I. Brugere, S. Boriah, V. Kumar, M. Steinbach, C. Potter, and S. Klooster. Incorporating natural variation into time-series based land cover

change identification. In *Proceeding of the 2011 NASA Conference on Intelligent Data Understanding (CIDU)*, 2011.

[180] V. Mithal, A. Khandelwal, S. Boriah, K. Steinhauser, and V. Kumar. Change detection from temporal sequences of class labels: Application to land cover change mapping. In *SIAM International Conference on Data mining, SDM. SIAM*, 2013.

[181] M. Montemerlo, S. Thrun, D. Koller, B. Wegbreit, et al. Fastslam 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges. In *International Joint Conference on Artificial Intelligence*, volume 18, pages 1151–1156. LAWRENCE ERLBAUM ASSOCIATES LTD, 2003.

[182] D. Neill, A. Moore, and G. Cooper. A bayesian spatial scan statistic. *Advances in neural information processing systems*, 18:1003, 2006.

[183] D. B. Neill, A. W. Moore, M. Sabhnani, and K. Daniel. Detection of emerging space-time clusters. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 218–227. ACM, 2005.

[184] J. Nieto, J. Guivant, E. Nebot, and S. Thrun. Real time data association for fastslam. In *Robotics and Automation, 2003. Proceedings. ICRA'03. IEEE International Conference on*, volume 1, pages 412–418. IEEE, 2003.

[185] D. Nolan and E. Rappin. Increased sensitivity of tropical cyclogenesis to wind shear in higher sst environments. *Geophys. Res. Lett*, 35, 2008.

[186] J. Nyberg, B. Malmgren, a. Winter, M. Jury, K. Kilbourne, and T. Quinn. Low atlantic hurricane activity in the 1970s and 1980s compared to the past 270 years. *Nature*, 447(7145):698, 2007.

[187] S. Oh, S. Russell, and S. Sastry. Markov chain monte carlo data association for multi-target tracking. *Automatic Control, IEEE Transactions on*, 54(3):481–497, 2009.

[188] K. Oouchi, J. Yoshimura, H. Yoshimura, R. Mizuta, S. Kusunoki, and a. Noda.

Tropical cyclone climatology in a global-warming climate as simulated in a 20 km-mesh global atmospheric model: Frequency and wind intensity analyses. *Journal of the Meteorological Society of Japan. Ser. Ii*, 84(2):259–276, 2006.

[189] J. Overpeck, G. Meehl, S. Bony, and D. Easterling. Climate data challenges in the 21st century. *Science*, 331(6018):700, 2011.

[190] B. Owens and C. Landsea. Assessing the skill of operational atlantic seasonal tropical cyclone forecasts. *Weather and Forecasting*, 18(1):45–54, 2003.

[191] E. PalmÈN. On the formation and structure of tropical hurricanes. *Geophysica*, 3:26–38, 1948.

[192] M. Paluš, D. Hartman, J. Hlinka, and M. Vejmelka. Discerning connectivity from dynamics in climate networks. *Nonlinear Processes Geophys.*, 18, 2011.

[193] W. Pegau, E. Boss, and A. Martínez. Ocean color observations of eddies during the summer in the gulf of california. *Geophysical Research Letters*, 29(9):1295, 2002.

[194] T. Perrone and P. Lowe. A statistically derived prediction procedure for tropical storm formation. *Monthly Weather Review*, 114(1):165–177, 1986.

[195] R. Pielke Jr, C. Landsea, M. Mayfield, J. Laver, and R. Pasch. Hurricanes and Global Warming. *Bulletin of the American Meteorological Society*, 86(11):1571–1575, 2005. ISSN 0003-0007.

[196] A. Poore. Multidimensional assignment and multitarget tracking. *Partitioning Data Sets. DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 19:169–196, 1995.

[197] J. Price. Upper ocean response to a hurricane. *Journal of Physical Oceanography*, 11:153–175, 1981.

[198] C. Race, M. Steinbach, A. Ganguly, F. Semazzi, and V. Kumar. a knowledge discovery strategy for relating sea surface temperatures to frequencies of tropical storms and generating predictions of hurricanes under 21st-century global warming scenarios. In *Cidu*, pages 204–212, 2010.

[199] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh. Searching and mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 262–270. ACM, 2012.

[200] R. Ramachandran, J. Rushing, H. Conover, S. Graves, and K. Keiser. Flexible framework for mining meteorological data. In *Proceedings of the 19th Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*, 2003.

[201] E. Rappaport, J. Franklin, L. Avila, S. Baig, J. Beven, E. Blake, C. Burr, J. Jiing, C. Juckins, R. Knabb, and Others. Advances and challenges at the national hurricane center. *Weather and Forecasting*, 24(2):395–419, 2009.

[202] D. Reid. An algorithm for tracking multiple targets. *Automatic Control, IEEE Transactions on*, 24(6):843–854, 1979.

[203] R. Reynolds, N. Rayner, T. Smith, D. Stokes, and W. Wang. An improved in situ and satellite sst analysis for climate. *Journal of Climate*, 15(13):1609–1625, 2002.

[204] P. Richardson. Eddy kinetic energy in the north atlantic from surface drifters. *Journal of Geophysical Research*, 88(C7):4355–4367, 1983.

[205] T. Rossby, M. Prater, and H. Søiland. Pathways of inflow and dispersion of warm waters in the nordic seas. *Journal of Geophysical Research*, 114(C4):C04011, 2009.

[206] S. M. Sall, H. Sauvageot, a. T Gaye, a. Viltard, and P. D. Felice. a cyclogenesis index for tropical atlantic off the african coasts. *Atmospheric Research*, 79(2): 123147, 2006. ISSN 0169-8095.

[207] B. D. Santer, T. M. L. Wigley, P. J. Gleckler, C. Bonfils, M. F. Wehner, K. Achutarao, T. P. Barnett, J. S. Boyle, W. Bruggemann, M. Fiorino, N. Gillett, J. E. Hansen, P. D. Jones, S. a. Klein, G. a. Meehl, S. C. B. Raper, R. W. Reynolds, K. E. Taylor, and W. M. Washington. Forced and Unforced Ocean Temperature Changes in Atlantic and Pacific Tropical Cyclogenesis Regions. *Proceedings of the*

National Academy of Sciences, 103(38):13905–13910, 2006. doi: 10.1073/Pnas.0602861103. URL `Http://Www.Pnas.Org/Content/103/38/13905.Abstract`.

[208] M. Saunders and a. Lea. Large contribution of sea surface warming to recent increase in atlantic hurricane activity. *Nature*, 451(7178):557–560, 2008.

[209] M. Scheffer, S. Carpenter, J. A. Foley, C. Folke, B. Walker, et al. Catastrophic shifts in ecosystems. *Nature*, 413(6856):591–596, 2001.

[210] a. Schumacher, M. Demaria, and J. Knaff. Objective estimation of the 24-h probability of tropical cyclone formation. *Weather and Forecasting*, 24(2):456–471, 2009.

[211] H. Sencan, Z. Chen, W. Hendrix, T. Pansombut, F. H. M. Semazzi, A. N. Choudhary, V. Kumar, A. V. Melechko, and N. F. Samatova. Classification of emerging extreme event tracks in multivariate spatio-temporal physical systems using dynamic network structures: Application to hurricane track prediction. In *IJCAI*, pages 1478–1484, 2011.

[212] Y. Serra, G. Kiladis, and M. Cronin. Horizontal and vertical structure of easterly waves in the Pacific Itcz. *Journal of the Atmospheric Sciences*, 65(4):1266–1284, 2008.

[213] L. Shapiro. Hurricane climatic fluctuations. part ii- relation to large-scale circulation. *Monthly Weather Review*, 110:1014–1023, 1982.

[214] L. Shapiro. Month-to-month variability of the atlantic tropical circulation and its relationship to tropical storm formation. *Monthly Weather Review*, 115:2598, 1987.

[215] S. Shekhar, R. R. Vatsavai, and M. Celik. Spatial and spatiotemporal data mining: Recent advances. *Data Mining: Next Generation Challenges and Future Directions*, 2008.

[216] I. Smal, K. Draegestein, N. Galjart, W. Niessen, and E. Meijering. Particle filtering

for multiple object tracking in dynamic fluorescence microscopy images: Application to microtubule growth analysis. *Medical Imaging, IEEE Transactions on*, 27 (6):789–804, 2008.

[217] D. Smith, R. Eade, N. Dunstone, D. Fereday, J. Murphy, H. Pohlmann, and a. Scaife. Skilful Multi-Year Predictions of Atlantic Hurricane Frequency. *Nature Geoscience*, 3:846–849, 2010. ISSN 1752-0894.

[218] R. Smith and P. Robinson. A bayesian approach to the modeling of spatial-temporal precipitation data. In *Case Studies in Bayesian Statistics*, pages 237–269. Springer, 1997.

[219] a. Solow and a. Beet. On the incompleteness of the historical record of north atlantic tropical cyclones (doi 10.1029/2008gl033546). *Geophysical Research Letters*, 35(11), 2008.

[220] a. Solow and L. Moore. Testing for a trend in a partially incomplete hurricane record. *Journal of Climate*, 13(20):3696–3699, 2000.

[221] R. Srikanthan, T. McMahon, et al. Stochastic generation of annual, monthly and daily climate data: A review. *Hydrology and Earth System Sciences Discussions*, 5(4):653–670, 2001.

[222] R. Sriver and M. Huber. Observational evidence for an ocean heat pump induced by tropical cyclones. *Nature*, 447(7144):577–580, 2007.

[223] M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, and C. Potter. Discovery of climate indices using clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 446–455. ACM, 2003.

[224] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly. Complex networks in climate science: progress, opportunities and challenges. In *NASA Conf. on Intelligent Data Understanding, Mountain View, CA*, 2010.

[225] P. Stolorz and C. Dean. Quakefinder: A scalable data mining system for detecting

earthquakes from space. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 208–213, 1996.

[226] P. Stolorz, E. Mesrobian, R. Muntz, J. Santos, E. Shek, J. Yi, C. Mechoso, and J. Farrara. Fast spatio-temporal data mining from large geophysical datasets. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 300–305, 1995.

[227] M. Sugi, a. Noda, and N. Sato. Influence of the global warming on tropical cyclone climatology: an experiment with the jma global model. *Journal of the Meteorological Society of Japan*, 80(2):249–272, 2002.

[228] G. Sugihara and R. May. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, 344(19):734–741, 1990.

[229] B. Tang and J. Neelin. Enso influence on atlantic hurricanes via tropospheric warming. *Geophys. Res. Lett*, 31:L24204, 2004.

[230] H. Taubenböck, T. Esch, A. Felbier, M. Wiesner, A. Roth, and S. Dech. Monitoring urbanization in mega cities from space. *Remote Sensing of Environment*, 2011.

[231] C. W. Team. *Climate Change 2007: Synthesis Report. Contribution of Working Groups I, Ii and Iii to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Ipcc, Geneva, Switzerland, 2007.

[232] C. Thorncroft and K. Hodges. African easterly wave variability and its relationship to atlantic tropical cyclone activity. *Journal of Climate*, 14(6):1166–1179, 2001.

[233] W. R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46:234–240, 1970.

[234] K. Trenberth. The Definition of El Niño. *Bulletin of the American Meteorological Society*, 78(12):2771–2777, 1997.

[235] K. Trenberth. Uncertainty in Hurricanes and Global Warming. *Science*, 308 (5729):1753, 2005.

[236] K. E. Trenberth and D. J. Shea. Atlantic hurricanes and natural variability in 2005. *Geophysical Research Letters*, 33:L12704, 2006.

[237] A. Tsonis and P. Roebber. The architecture of the climate network. *Physica A: Statistical Mechanics and its Applications*, 333:497–504, 2004.

[238] A. A. Tsonis, K. L. Swanson, and P. J. Roebber. What do networks have to do with climate? *Bulletin of the American Meteorological Society*, 87(5):585–596, 2006.

[239] A. A. Tsonis, K. L. Swanson, and G. Wang. On the role of atmospheric teleconnections in climate. *Journal of Climate*, 21(12):2990–3001, 2008.

[240] U. Ulbrich, G. Leckebusch, and J. Pinto. Extra-tropical cyclones in the present and future climate: a review. *Theoretical and Applied Climatology*, 96(1):117–131, 2009.

[241] T. T. van Leeuwen, A. J. Frank, Y. Jin, P. Smyth, M. L. Goulden, G. R. van der Werf, and J. T. Randerson. Optimal use of land surface temperature data to detect changes in tropical forest cover. *Journal of Geophysical Research*, 116(G2): G02002, 2011.

[242] G. Vecchi and B. Soden. Increased tropical atlantic wind shear in model projections of global warming. *Geophys. Res. Lett*, 34, 2007.

[243] G. Vecchi and B. Soden. Effect of remote sea surface temperature change on tropical cyclone potential intensity. *Nature*, 450(7172):1066–1070, 2007.

[244] G. Vecchi, M. Zhao, H. Wang, G. Villarini, a. Rosati, a. Kumar, I. Held, and R. Gudgel. Statistical–dynamical predictions of seasonal north atlantic hurricane activity. *Monthly Weather Review*, 139:1070–1082, 2011.

[245] G. a. Vecchi and T. R. Knutson. on estimates of historical north atlantic tropical cyclone activity. *Journal of Climate*, 21(14):3580–3600, 2008. doi: 10.1175/2008Jcli2178.1. URL `Http://Journals.Ametsoc.Org/Doi/Abs/10.1175/2008Jcli2178.1`.

[246] F. Vitart. Seasonal forecasting of tropical storm frequency using a multi-model ensemble. *Qjr Meteorol. Soc*, 132:647–666, 2006.

[247] F. Vitart, J. Anderson, and W. Stern. Simulation of interannual variability of tropical storm frequency in an ensemble of gcm integrations. *Journal of Climate*, 10(4):745–760, 1997.

[248] F. Vitart, M. Huddleston, M. DÉQuÉ, D. Peake, T. Palmer, T. Stockdale, M. Davey, S. Ineson, and a. Weisheimer. Dynamically-based seasonal forecasts of atlantic tropical storm activity issued in june by eurosip. *Geophysical Research Letters*, 34(16):L16815, 2007.

[249] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2): 1–305, 2008.

[250] G. Walker and E. Bliss. World weather iv: Some application to seasonal foreshadowing. *Memoirs of the Royal Meteorological Society*, 3(24):81–95, 1930.

[251] D. Watts and S. Strogatz. The small world problem. *Collective Dynamics of Small-World Networks*, 393:440–442, 1998.

[252] P. Webster and H. Chang. Atmospheric wave propagation in heterogeneous flow: Basic flow controls on tropical–extratropical interaction and equatorial wave modification. *Dynamics of Atmospheres and Oceans*, 27(1-4):91–134, 1998.

[253] P. Webster and S. Yang. Monsoon and enso: Selectively interactive systems. *Quarterly Journal of the Royal Meteorological Society*, 118(507):877–926, 1992.

[254] P. Webster, J. Curry, J. Liu, and G. Holland. Response to comment on" changes in tropical cyclone number, duration, and intensity in a warming environment". *Science*, 311(5768):1713, 2006.

[255] P. J. Webster, G. J. Holland, J. a. Curry, and H. Chang. Changes in tropical cyclone number, duration, and intensity in a warming environment. *Science*, 309(5742):1844 –1846, 2005. doi: 10.1126/Science.1116448. URL `Http://Www.Sciencemag.Org/Content/309/5742/1844.Abstract`.

[256] M. A. White, F. Hoffman, W. W. Hargrove, and R. R. Nemani. A global framework for monitoring phenological responses to climate change. *Geophysical Research Letters*, 32(4):L04705, 2005.

[257] D. S. Wilks. *Statistical methods in the atmospheric sciences.* Academic press, 2006.

[258] K. Wolter and M. Timlin. El niño/southern oscillation behaviour since 1871 as diagnosed in an extended multivariate enso index (mei. ext). *International Journal of Climatology.*

[259] D. A. Woolhiser. Modeling daily precipitation progress and problems. In W. A and G. P, editors, *Statistics in the Environmental and Earth Sciences.* Edward Arnold, London, 1992.

[260] E. Wu, W. Liu, and S. Chawla. Spatio-temporal outlier detection in precipitation data. In *Proceedings of the Second international conference on Knowledge Discovery from Sensor Data*, pages 115–133. Springer-Verlag, 2008.

[261] E. Wu, W. Liu, and S. Chawla. Spatio-temporal outlier detection in precipitation data. *Knowledge discovery from sensor data*, pages 115–133, 2010.

[262] K. Wyrtki, L. Magaard, and J. Hager. Eddy energy in the oceans. *Journal of Geophysical Research*, 81(15):2641–2646, 1976.

[263] K. Yamasaki, A. Gozolchiani, and S. Havlin. Climate networks around the globe are significantly affected by el nino. *Physical review letters*, 100(22):228501, 2008.

[264] S. Yeh, J. Kug, B. Dewitte, M. Kwon, B. Kirtman, and F. Jin. El niño in a changing climate. *Nature*, 461(7263):511–514, 2009.

[265] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys (CSUR)*, 38(4):13, 2006.

[266] J. Yoshimuea, M. Sugi, and a. Noda. Influence of greenhouse warming on tropical cyclone frequency. *Journal of the Meteorological Society of Japan*, 84(2):405–428, 2006.

[267] R. Zhang and T. Delworth. Impact of atlantic multidecadal oscillations on india/sahel rainfall and atlantic hurricanes. *Geophys. Res. Lett*, 33:L17712, 2006.

[268] M. Zhao, I. Held, S. Lin, and G. Vecchi. Simulations of global hurricane climatology, interannual variability, and response to global warming using a 50-km resolution gcm. *Journal of Climate*, 22(24):6653–6678, 2009.

Appendix A

List of Symbols and Acronyms

Table A.1: Acronyms (adopted from [118])

Acronym	Meaning
AAO	Antarctic Oscillation.
AO	Arctic Oscillation - The leading empirical orthogonal function (EOF) of sea level pressure poleward of 20°N
Coriolis parameter	A measure that is twice the local vertical component of the angular velocity of a spherical planet, 2Ω sinφ, where Ω is the angular speed of the planet and φ is the latitude.
CLIPER	Trend and 2- and 7-yr variations of the TC counts
Climatology	Long-term mean of a seasonal activity.
ENSO	El NiñoSouthern Oscillation.
H	Hurricane - A tropical cyclone with sustained low level winds of 74 mph (33 m s1 or 64 kt) or greater.
HC	Index of the frequency of clod-air September-December and January-May
HD	Hurricane day - A measure of hurricane activity, one unit of which occurs as four 6-h periods during which a tropical cyclone is observed or estimated to have hurricane intensity winds.

Continued on next page

Table A.1 – continued from previous page

Acronym	Meaning
HIB	Index of the strength of the India-Burma through (15-20N, 80-100E) at 500 hPa
HPV	Index of the area of the polar vortex in the Pacific sector (150E-120W)
HSCS	Index of the northern extent of the subtropical high over the South China Sea (100-120E)
HTP	Index of the strength of the subtropical high over Tibet (25-35N, 80-100E) at 500 hPa
HWNP	Index of the Westward extent of the 5880-m contour of the 500-hPa subtropical high over the western North Pacific
IH	Intense hurricane - A hurricane that reaches a sustained 1-min-average 10-m-height wind of at least 111 mph (96 kt or 50 m s1) at some point in its lifetime. This constitutes a category 3 or higher on the Saffir-Simpson scale (also termed a major hurricane).
IHD	Intense hurricane day - Four 6-h periods during which a hurricane has intensity of SaffirSimpson category 3 or higher.
MJO	Madden-Julian oscillation.
NAO	North Atlantic OscillationA normalized measure of the surface pressure difference between Iceland and Portugal.
NS	Named storm - A hurricane or tropical storm.
NSD	Named storm day - As in HD but for four 6-h periods during which a tropical cyclone is observed (or is estimated) to have attained tropical storm-intensity winds.
NTC	Net tropical cyclone activity - Average seasonal percentage mean of named storms, named storm days, hurricanes, hurricane days, intense hurricanes, and intense hurricane days. Gives overall indication of Atlantic basin seasonal hurricane activity.
NPO	North Pacific Oscillation.

Continued on next page

Table A.1 – continued from previous page

Acronym	Meaning
PNA	Pacific-North American pattern - One of the most prominent modes in low-frequency variability in the extratropics of the Northern Hemisphere, appearing in all months except June and July.
PWAT	Precipitable water.
QBO	Quasi-biennial oscillation - A stratospheric (16-35-km altitude) oscillation of equatorial east west winds that vary with a period of about 26-30 months or roughly 2 yr, typically blowing for 12-16 months from the east, then reversing and blowing 12-16 months from the west, then back to east again.
SLP	Sea level pressure.
SLPA1	Southern Hemisphere sea level pressure anomaly 1 - mean sea level pressure anomaly averaged over the region (5-35S, 180-150W)
SLPA2	Southern Hemisphere sea level pressure anomaly 2 - mean sea level pressure anomaly averaged over the region (5-35, 150120W).
SST	Sea surface temperature.
SSTA	Sea surface temperature anomaly.
SHO	Siberian High Oscillation.
TC	Tropical cyclone - A large-scale circular flow occurring within the Tropics and subtropics that has its strongest winds at low levels, including hurricanes, tropical storms, and other weaker rotating vortices.
TOH	Tropical-only hurricanes.
TONS	Tropical-only named storms.
TS	Tropical stormA tropical cyclone with maximum sustained winds between 39 mph (18 m s1 or 34 kt) and 73 mph (32 m s1 or 63 kt).
V850	Australian monsoon circulation index - 850 hPa meridional winds averaged over the region (5-20S, 140-170E),

Continued on next page

Table A.1 – continued from previous page

Acronym	Meaning
Vor	Relative vorticity.
VWS	Vertical wind shear.
WP	West Pacific Pattern Index.

Appendix B

Review of TC Statistical Forecast Models

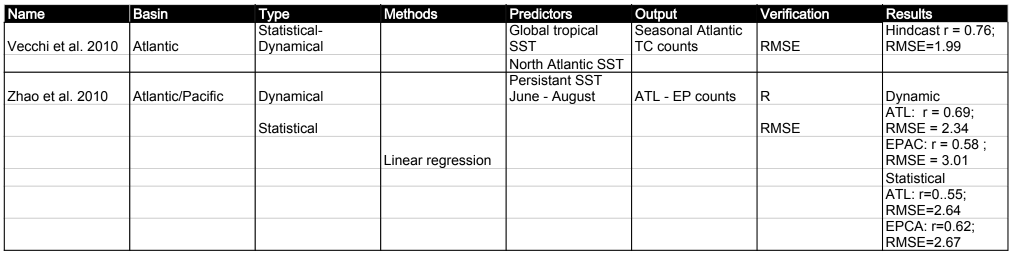
B.1 Summary of Statistical Modeling Studies

B.2 Summary of Testing Methods

Name	Basin	Type	Methods	Predictors	Output	Verification	Results
Belanger et al. 2010	Atlantic	Dynamical				BSS	
						ROC	
Belanger et al. 2010	North Indian	Dynamical				BSS	0.17 - 0.82 (Arabian Sea)
						ROC	0.09 - 0.80 (Bay of Bengal)
Blake and Gray 2004	Atlantic	Statistical		Jul 200-mb v wind (4S-8N 105-79W)	August NTC	Jackknife	RMSE = 5.17
				Jul SLPA (25–37.5 N, 47.5–25W)			
				Jul SLPA (47–62N, 156E–164W)			
				Jul 200-mb u wind (40–35S, 110–85W)			
				Jul 500-mb height (42.5–27.5S, 72.5–95E)			
				Jul 200-mb u wind (17.5-7.5S, 145-180E)			
				Jul 200-mb u wind (5S–5N, 110–85W)			
				Jun 200-mb u wind (80–85N, 45W–10E)			
				Jun SLPA (18–30N, 134–154E)			
				Apr SLPA (10S–5N, 35W–15E)			
				Feb SLPA (52.5–75N, 5W–35E)			
				Jan SLPA (30–40N, 110-95W)			
Chan et al. 1998	North Pacific	Satitistical	Smooth Multiple Additive Regression Technique (SMART)	SOI	Annual number of TCs (TCA), ‘‘A’’ meaning annual	RMSE, R, Measure of Agreement	TCA: R=0.89; RMSE=2.9; MA = 0.56
	South China Sea			NINO4	Annual number of tropical storms and typhoons (TSYA)		TSYA: R=0.86; RMSE=2.6; MA=0.48
				WP	Annual number of typhoons (TYA)		TYA: R=0.75; RMSE=2.5; MA=0.43
				HSCS	Number of tropical storms and typhoons for 8 months (May - December) (TSY8)		TSY8: R=0.86; RMSE=2.4; MA = 0.53
				HWNP	Number of typhoons (TY8).		TC8: R=0.72; RMSE=2.6; MA=0.36
				HPV	Annual number of TCs (TCS), ‘‘S’’ meaning South China Sea		TCS: R=0.77; RMSE=2.7; MA=0.34
				HTP	Annual number of tropical storms and typhoons (TSYS).		TSYS: R=0.75; RMSE=1.9; MA=0.38
				HIB			
				HC			
				CLIPER			
Chan et al. 2001	North Pacific	Satitistical	Smooth Multiple Additive Regression Technique (SMART)	SOI	Annual number of TCs (TCA), ‘‘A’’ meaning annual	RMSE, R, Measure of Agreement	TCA: R=0.84; RMSE=3.4; MA = 0.51
	South China Sea			NINO4	Annual number of tropical storms and typhoons (TSYA)		TSYA: R=0.73; RMSE=3.9; MA=0.36
				WP	Annual number of typhoons (TYA)		TYA: R=0.57; RMSE=3.2; MA=0.33
				HSCS	Annual number of TCs (TCS), ‘‘S’’ meaning South China Sea		TCS: R=0.71; RMSE=2.8; MA=0.33
				HWNP	Annual number of tropical storms and typhoons (TSYS).		TSYS: R=0.69; RMSE=2.0; MA=0.35
				HPV			
				HTP			
				HIB			
				HC			
				CLIPER			
				V850			
				SLPA1			
				SLPA2			
Choi et al. 2010	Western North Pacific	Statistical	Multiple Linear Regression Model	April-May SHO	TC genesis frequency	Cross validation, R	R= 0.65, RMSE = 2.31
				April-May AAO			
				April-May NPO			

Name	Basin	Type	Methods	Predictors	Output	Verification	Result
Chu and Zhao 2007	Central North Pacific	Statistical	Bayesian Regression with Poisson prior	SST	Annual TC counts	Cross validation	N/A
				850 hPa Vor			
				PWAT			
				1000- 850- 200 hPa SLP			
				VWS			
Chu et al. 2007	Taiwan	Statitical	LAD regression	SST		R, cross validation	
				850 hPa Vor			
				PWAT			
				850- 250 hPa SLP			
				VWS			
Elsner and Jagger (2006)	Atlantic		Bayesian with counts pre-1899 treated as less certain	1. May - June NAO	Seasonal Hurricane coastal activity	Leave-one-out cross validation	
				2.May - June SOI			
				3.May - June AMO			
				4. Jul - Dec Hurricane counts			
Fan 2010	Atlantic	Statistical	Linear regression	200-hPa air temperature in Eurasia (308–458N, 608–1058E) in November of the previous year	Year-to-year TC counts	MAE (mean average absolute error)	29% improvement over climatological MAE for hindcast (2004-2009)
				Air temperature at 200 hPa in the Atlantic (308–408N, 458–158W) in November of the previous year		Cross validation	46% improvement over climatological MAE for cross validation (1985-2009)
				Air temperature at 200-hPa south of 608S in February of the current year			
				SLP over the South Pacific (208–58S, 1208–908W) in March of the current year			
				200-hPa geopotential height over the northwest Pacific (108–208N, 1508–1658E) in April of the current year			
				200-hPa zonal wind over the North Pacific (458–658N, 1608–1208W) in April of the current year			
Gonzalez et al. 2011	North Atlantic		Hierarchical classifier-regressor model	ATL: Global Annual SLP	Seasonal TC counts for Atlantic and North Pacific	10-Fold corss validation	NA-NP
	North Pacific		Least-Absolute Deviation (LAD) regression	Global Annual SST		Leave one out cross validation	
				Global Annual vertical wind shear		RMSE	0.77-0.99
				Precipitable water		R	0.87-091
				NP:		R^2	0.76-0.82
				0-30N? SLP, PW, SST, VWS		MSSS	0.79-0.77
						GSS	0.92-0.92
						HSS	0.95-0.94
Gray et al. (1992)	Atlantic		Forward extrapolation of QBO	QBO (Sep-Nov)	1. Number of named TCs	Cross-validation (jackknife)	Correlation of 0.44 - 0.51 between forecast and actual
			Least absolute deviation (LAD) regression	Western Sahel rainfall (Aug-Sept)	2. Named of named TC days	Agreement of cross-validation results under null hypothsis	
				Gulf of Guinea rainfall (Aug-November)	3. Number of hurricanes		
					4. Number of hurricane days		
					5. Number of major hurricanes		
					6. Named of major hurricane days		
					7. Hurricane destruction potential (HDP)		

Name	Basin	Type	Methods	Predictors	Output	Verification	Results
Klotzbach and Gray 2004	Seasonal		Poisson regression for TCs	November 500-mb geopotential height in the far North Atlantic (67.5°–85° N, 10°E–50°W)	Net tropical cyclone (NTC) activity prediction	Jackknife cross-validation as in Elsner and Schmertmann (1993)	
				October–November SLP in the Gulf of Alaska (45°–65°N, 120° –160°W)			
				September 500-mb geopotential height in western North America (35° –55°N, 100°–120° W)			
				July 50-mb equatorial zonal wind (5°S–5°N, all longitudes)			
				September–November SLP in the Gulf of Mexico–southeastern United States (15° –35°N, 75°–95°W)			
				November SLP in the tropical northeast Pacific (7.5°–22.5°N, 125° –175°W)			
Klotzbach and Gray 2003	September counts		Least square linear regression	April 1000-mb U (12.5°–30°S, 40° W–10°E)			
			top 10–bottom 10 composite maps of various meteorological parameters such as sea surface temperature, sea level pressure, and zonal wind at 200, 850, and 1000 mb for the month of September by differencing active and inactive Septembers	July 200-mb geopotential height (32°–42°N, 100° –160°E)			
			NOAA correlation tool	July–August 1000-mb U (5°–15°N, 30°–50°W)–(22.5° –35°N, 35°–65°W)			
				February 1000-mb U (20°–30°N, 15° W–15°E)			
				April 200-mb U (67.5°–85°N, 110° –180°E) (			
				August SLP (0° –30°S, 120°–160° E)			
				August SLP (20° –45°S, 60°–90°E)			
				May 200-mb V (0° –20°S, 15°–30°E)			
				January–February 200-mb U (15° –25°N, 120°E–160°W)			
Lu et al. 2010	Taiwan	Statistical	Bayesian Regression with Poisson prior	SST		HSS, HK, Accuracy	Accuracy = 0.52, HK=0.27; HSS = 0.27
			Stepwise regression for predictor selection	850 hPa Vor			
				PWAT			
				SLP			
				VWS			
Nicholls 1992	Australian region (105 E- 165E)	Statistical	Linear regression	September-November mean SOI	Seasonal TC counts	RMSE	RMSE = 2.79
			Predict TC count then add previous year's count				RMSE_climatology = 3.71
Nyberg et al. 2007	Atlantic		Neural Network	Aug-Oct MDR SST (ERSST v2)	Aug-Oct vertical wind shear (|V_z| m s^{-1}) centered at 11N and 65W	Root mean square error	.97 correlation with overlapping reconstructed and instrumental periods (1949-1990)
			Paleoclimatic records	Agu-Oct luminescence intensity from coral cores	Aug-Oct major hurriccanes		
				G. bulloids concentration			



Name	Basin	Type	Methods	Predictors	Output	Verification	Results
Vecchi et al. 2010	Atlantic	Statistical-Dynamical		Global tropical SST	Seasonal Atlantic TC counts	RMSE	Hindcast r = 0.76; RMSE=1.99
				North Atlantic SST			
Zhao et al. 2010	Atlantic/Pacific	Dynamical		Persistant SST June - August	ATL - EP counts	R	Dynamic
		Statistical				RMSE	ATL: r = 0.69; RMSE = 2.34
			Linear regression				EPAC: r = 0.58 ; RMSE = 3.01
							Statistical
							ATL: r=0..55; RMSE=2.64
							EPCA: r=0.62; RMSE=2.67

Table B.1: Forecast evaluation and verification methods. Adapted from [28]

Method	Description	Studies
Root mean square error (RMSE)	Over several runs, what's the difference between forecasts and observations. Target value: 0	Nyberg *et al.* 2007[186], Vecchi *et al.* 2011 [244]
Pearson Correlation (R)	The magnitude of forecast error. Target value: 1	
Spearman rank correlation coefficient (R_s)	Nonparametric correlation. Target value: +/-1	Nyberg *et al.* 2007[186]
Uncentred correlation coefficient	For shorter time-series. The mean is not substracted when computing standard deviations. Target value: 1	
Anomaly correlation coefficient (ACC)	The correlation between forecasts and deviations from climatology.	Smith 2010 [217]
Ranked probability skill score (RPSS)	Improvement of probability forecast over climatology. Target value: 1	
Brier skill score (BSS)	Relative skill of forecasts over climatology for binary events. Target value: 1	Belanger *et al.* 2010 [17], Belanger *et al.* 2011 [18]

Continued on next page

Table B.1 – continued from previous page

Method	Description	Studies
Relative operating characteristic (ROC)	The model's ability to discriminate between positive and negative events. Target value: 1	Belanger *et al.* 2010 [17], Belanger *et al.* Belanger 2011 [18]
Persistence	The model's ability to match the previous year's activity	Smith 2010 [217], Vecchi *et al.* 2011 [244]
Heidke skill score (HSS)	The fraction of correct classifications minus forecasts that would be correct due to random chance. Target value: 1	
Cross validation (Jackknife)	How well does the model generalize for an independent dataset.	Chan *at al.* 1998 [40]
Gerrity Skill Score (GSS)	A reward/penalty matrix for event prediction. Target score: 1	Gonzalez *et al.* 2011 [113]
Mean Square Skill Score (MSSS)	Forecast mean square error relative climatology mean square error. Target value: 1	Gonzalez *et al.* 2011 [113]