Spatio-Temporal Data Mining for Climate Data: Advances, Challenges, and Opportunities

James H. Faghmous and Vipin Kumar

Abstract Our planet is experiencing simultaneous changes in global population, urbanization, and climate. These changes, along with the rapid growth of climate data and increasing popularity of data mining techniques may lead to the conclusion that the time is ripe for data mining to spur major innovations in climate science. However, climate data bring forth unique challenges that are unfamiliar to the traditional data mining literature, and unless they are addressed, data mining will not have the same powerful impact that it has had on fields such as biology or e-commerce. In this chapter, we refer to spatio-temporal data mining (STDM) as 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). We highlight some of the singular characteristics and challenges STDM faces within climate data and their applications, and provide the reader with an overview of the advances in STDM and related climate applications. We also demonstrate some of the concepts introduced in the chapter's earlier sections with a real-world STDM pattern mining application to identify mesoscale ocean eddies from satellite data. The case-study provides 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. We end the chapter with a discussion of notable opportunities for STDM research within climate.

James H. Faghmous
Department of Computer Science and Engineering, The University of Minnesota – Twin Cities
e-mail: jfagh@cs.umn.edu

Vipin Kumar
Department of Computer Science and Engineering, The University of Minnesota – Twin Cities
e-mail: kumar@cs.umn.edu

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 [39]. 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 [69].

The last decades have seen tremendous growth in data-driven learning algorithms and their broad-range applications [46]. 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 atmospherehydrosphereland 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 chapter 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). We begin by briefly reviewing the different types of climate data available and expand on notable challenges associated with them. We then proceed to a broad review of sample works in the STDM literature applied to climate spatio-temporal data. We then demonstrate the promise of STDM on a real-world application of tracking mesoscale ocean eddies in satellite data. We conclude the chapter with a review and future directions.

2 An Overview of Climate Data and Associated Challenges

In this section, we review the different types of climate data available to data mining researchers and the notable caveats when mining climate data.

2.1 Types of Climate Data

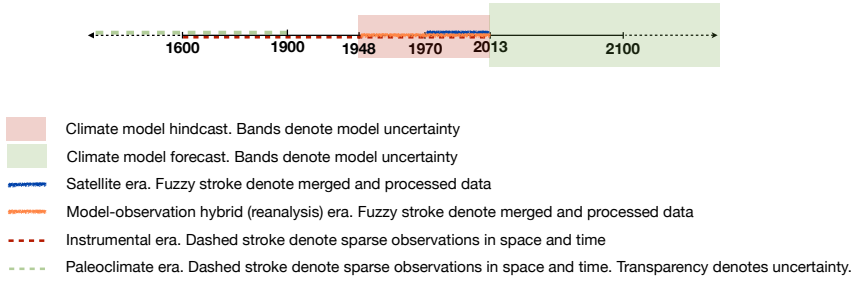

Fig. 1 Climate science has numerous types of data, each with its own challenges.

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 [69]. 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 [86]. 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 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 smoother than the raw observational records and have extended spatio-temporal coverage. While reanalyses are considered the

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

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 or temperature 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.

2.2 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) data 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 is more commonly known as *autocorrelation* dictates that spatio-temporal data not be modeled as statistically independent [87]. 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 [25]. 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.

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

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 challenge. 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. Given that each dataset has its own biases, it unclear what effect fusing these datasets would have on data mining tasks and knowledge extracted therein.

Additionally, 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 section, we will provide the reader with a concise review of the STDM literature pertaining to climate data.

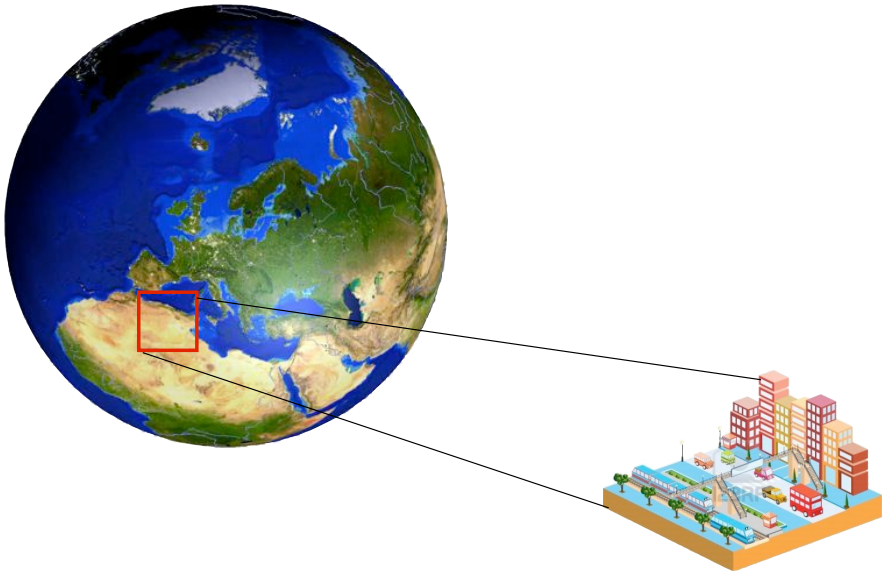

Fig. 2 A large amount of climate data is at at global spatial scale (∼250km), however many climate-related questions are at the regional (∼50km) or local (km or sum-km) scale. This multi-scale discrepancy is a significant data mining challenge.

3 Advances in STDM applications to Climate

Although the fields of temporal and spatial data mining research are relatively mature [77, 56], 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 [22, 21, 57, 27]. 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 subsections, we will pro-

vide 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.* [61] 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 [61, 83]. As en extension to CONQUEST, Stolorz and Dean [82] introduced Quakefinder, an automatic application that detects and measures tectonic activity from remote sensing imagery. Mesrobian *et al.* [62] introduced Oasis, an extensible exploratory data mining tool for geophysical data. A similar application is the algorithm development and mining framework (ADaM) [73] which was developed to mine geophysical events in spatio-temporal data. Finally, Baldocchi *et al.* [4] 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.* [36, 72]).

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 [25], 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 [63] analyzed the rotated[3] EOFs of global SST data and linked the first six principal components to ocean-atmospheric modes[4]. In another application, Basak *et al.* [6] used independent component analysis to discover the North Atlantic Oscillation index (NAO) [55] in SLP data. For a comprehensive discussion of EOF analysis for climate data please see [97].

Within clustering applications, Hoffman *et al.* [48] developed a spatio-temporal clustering algorithm to identify regions with similar environmental characteristics. White *et al.* [96] applied the techniques presented in [48] to generate climate and vegetation clusters that were subsequently used to infer phenological responses to climate change. Braverman and Fetzer [9] mined large-scale structures in climate data using a data compression technique based on entropy-constrained vector quantization [20] to generate multivariate distribution estimates of the data and monitored the changes of such distributions across space, time, and resolution. McGuire *et al.* [60] used spatial neighborhood and temporal discretization methods to identify spatio-temporal neighborhoods in SST data. In another clustering application, Gaffney *et al.* [42] clustered cyclone tracks using a regression mixture model and works by Camargo *et al.* [10] and Camargo *et al.* [11] 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 [91]. Another dominant climate pattern is the InterTropical Convergence Zone (ITCZ), a phenomena on a daily time scale over the east Pacific. Bain *et al.* [3] developed a spatio-temporal Markov random field to detect the ITCZ in satellite data. Henke *et al.* [47] extended such methods by using a super- and semi-supervised method to track this dynamic phenomena and its properties in satellite

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

[4] Emanuel [33] points out that EOFs are *not* mathematically equivalent to modes

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 3 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.

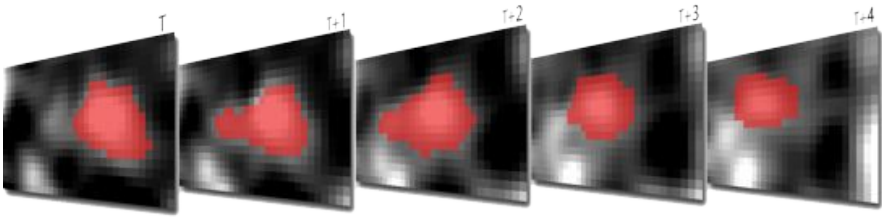

Fig. 3 An ocean eddy moving in time as detected in ocean data. One of the challenges of STDM is to identify significant patterns in continuous spatio-temporal climate data.

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, 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 4 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.* [8] 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.* [64] introduced a global land-cover

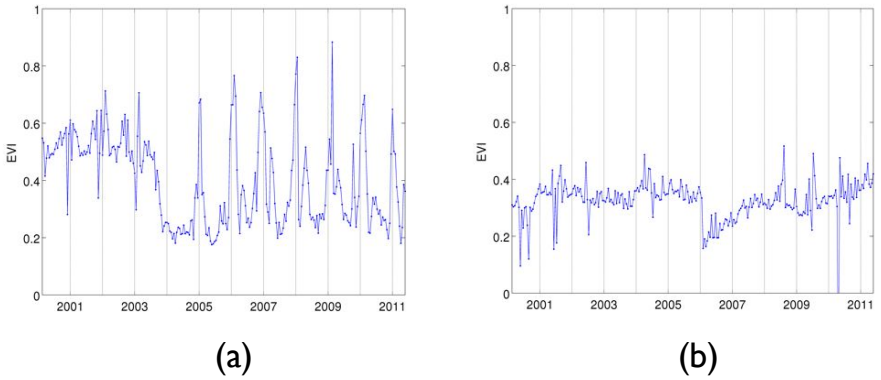

Fig. 4 An example of a spatio-temporal event (a) and anomaly (b). The time-series denote changes in vegetation over time. (a) A land-cover change event as seen in the decrease of vegetation due to agricultural expansion in 2003. (b) an abrupt drop in vegetation due to a forest fire in 2006, the vegetation gradually returned after the fire.

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.* [65] provide a concise discussion of STDM techniques and challenges related to land-cover change. In another global-scale event detection application, Fu *et al.* [41] extended the traditional Markov random field (MRF) model [93] 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 [68, 67] 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 [19] 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 [5] used a parallel wavelet transform to detect spatio-temporal outliers in SST data. Wu *et al.* [99, 100] 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 [1] 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 [13], 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.* [43] 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 [84] 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 [75]. 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.

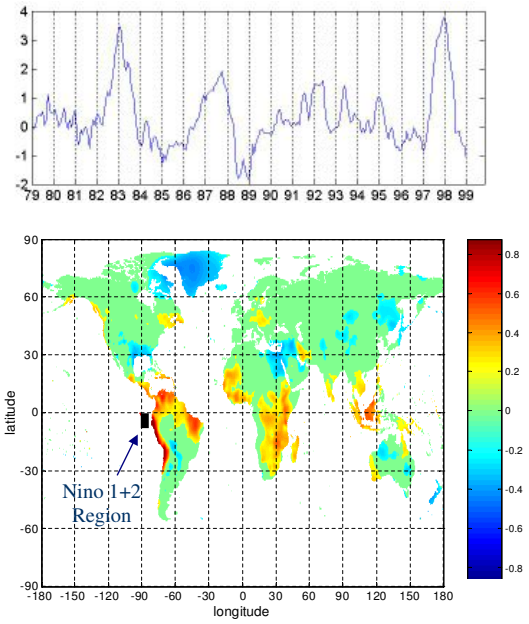

Fig. 5 Top: The NINO1+2 time-series which was constructed by averaging the sea surface temperatures (SST) of the box highlighted in the map below. Bottom: the linear correlation between the NINO1+2 index and global land surface temperature anomalies.

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 5 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 [44] used linear and partial linear correlations to link vertical wind shear in the Atlantic to SST and Sahel rainfall patterns. Webster *et al.* [95] 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.* [18] 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 [18] 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.* [49]), climate scientist use *composite analysis* as a another way to quantify how well one variable explains another. Figure 6 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 6'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 6 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 [54] 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 [31] to determine how significant was the mean of the small sample relative to random noise.

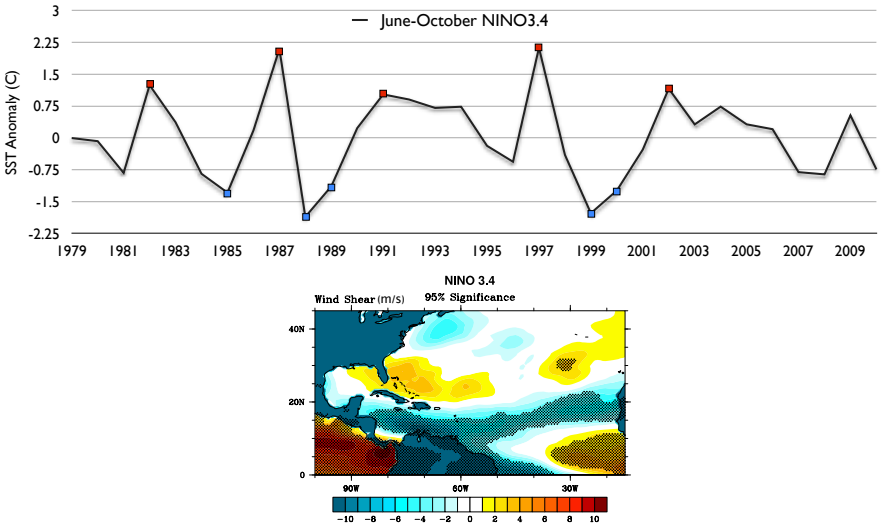

Fig. 6 An example on how composites un-earth non-linear relationships between variables. Top panel: time-series of SST anomalies in the NINO3.4 region. Bottom panel: Composite of June-October mean vertical wind shear, which was constructed by subtracting the top panel's mean of the negative extremes 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 7 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.

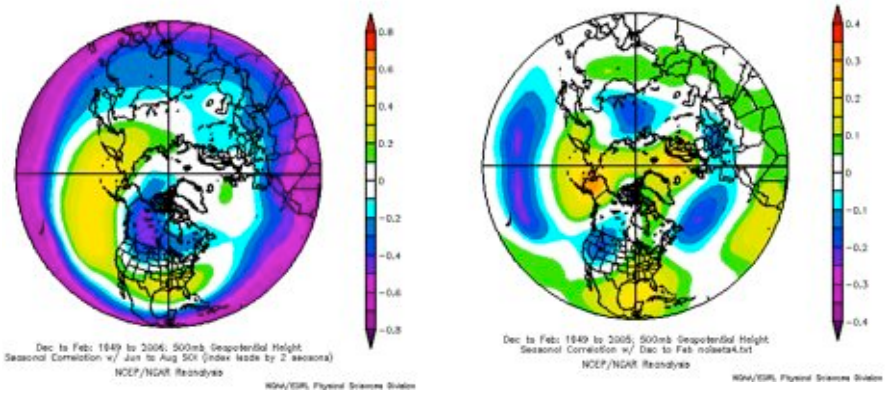

Fig. 7 Geopotential height correlated with the Southern Oscillation Index (SOI; left) and random noise (right). This is an example how high and spatially coherent correlations can be the result of random chance.

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 [23] 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 [24] 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 [98, 78, 79]. Huang and Cressie [50] 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.* [26] 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.* [92] built a logistic regression-based model trained on land surface temperatures to detect changes in tropical forest cover. Karpatne *et al.* [51] extended the work in [92] 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 [85]. 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.* [66] 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.* [14] 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 [88], mutual information [29], syntonization [2], *etc.*) Once a network is built, it is possible to apply the techniques previously discussed such as relationship mining [52], predictive modeling [81, 76], or pattern mining [80] on the transformed data.

Steinbach *et al.* [80] 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.* [52] extended the work in [80] to allow for dynamic dipoles (strongly correlated distant spatial regions) in climate data. Kawale *et al.* [53] proposed a bootstrapping method to test the significance of such long-range spatio-temporal patterns.

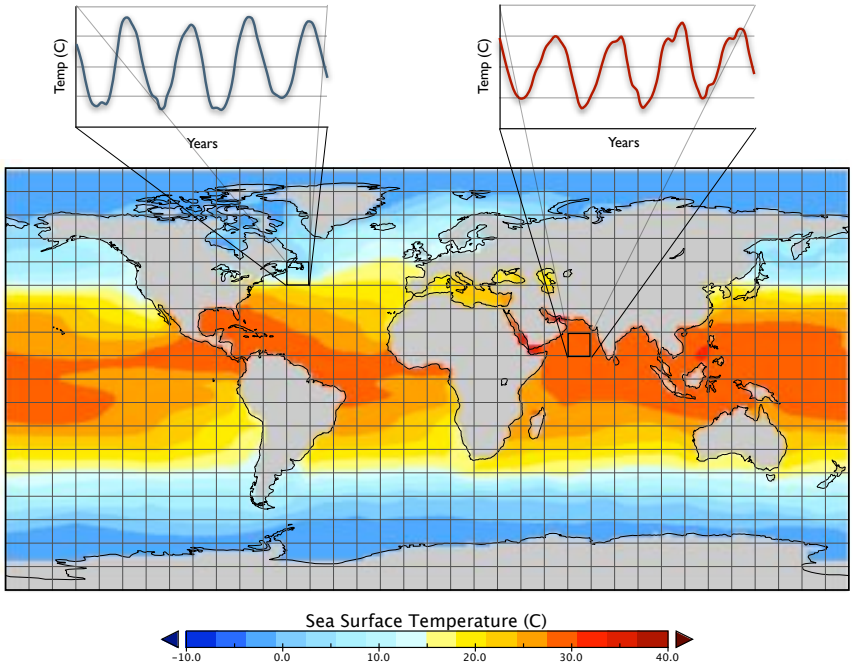

Fig. 8 Gridded spatio-temporal climate data can be analyzed in a network format. Each grid location is characterized by a time series. A network can be constructed between each location with an edge weight being the relationship between the time-series of each location.

Inspired by complex networks, [88] 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 [30, 102, 45]. Similarly, Tsonis *et al.* [89] showed that some climate phenomena and datasets obey a small-world network property [94]. Furthermore, several studies found distinct structural differences between the networks around tropical and extra-tropical regions [89, 29]. Berezin *et al.* [7] 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 [89, 90].

Others have built networks using non-gridded discrete climate data. Elsner *et al.* [32] used seasonal hurricane time-series to construct a network to study interannual hurricane count variability. Fogarty *et al.* [38] 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 [70].

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.

4 STDM Application Case Study: Ocean Eddies Monitoring

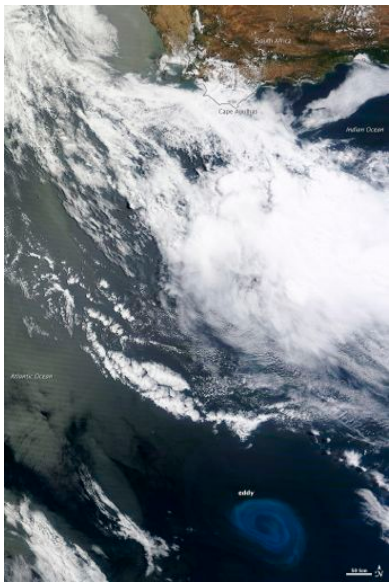

Fig. 9 Image from the NASA TERRA satellite showing an anti-cyclonic (counter-clockwise in the Southern Hemisphere) eddy that likely peeled off from the Agulhas Current, which flows along the southeastern coast of Africa and around the tip of South Africa. This eddy (roughly 200 km wide) is an example 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.

In this section, we will provide an in-depth case study for mining patterns in continuous climate data, highlight some of the challenges discussed in previous sections, and provide possible ways to address them.

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 [101, 74, 15]. 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 9). In contrast to linear Rossby waves, the rotation of nonlinear 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 10). 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.* [103] 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 [69].

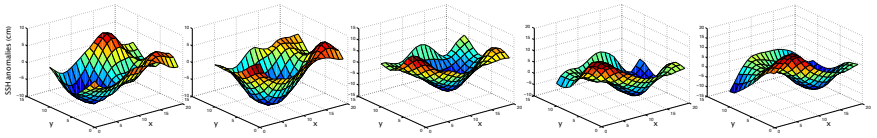

Fig. 10 An example of a cyclonic eddy traveling through a continuous sea surface height (SSH) field (from left to right). Unlike common feature mining and tracking tasks, features in physical sciences are often not self-defined with unambiguous contours and properties. Instead, they tend to be dynamic 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.

Our understanding of ocean eddy dynamics has grown significantly with the advent of satellite altimetery. Prior to then, oceanographers relied primarily on case studies using drifting floats in the open ocean to collect detailed information about individual eddies such as rotational speeds, amplitude, and salinity profiles. With the increased accessibility to satellite data, ocean surface temperatures and color have been used to identify ocean eddies based on their signatures on such fields [71, 37, 28]. While, these fields are impacted by eddy activity, there are additional

phenomena, such as hurricanes or near-surface winds, that affect them as well; effectively complicating eddy identification in such data fields. More recently, sea surface height (SSH) observations from satellite radar altimeters have emerged as a better-suited alternative for studying eddy dynamics on a global scale given SSH's intimate connection to ocean eddy activity. 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. In Figure 11, 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.

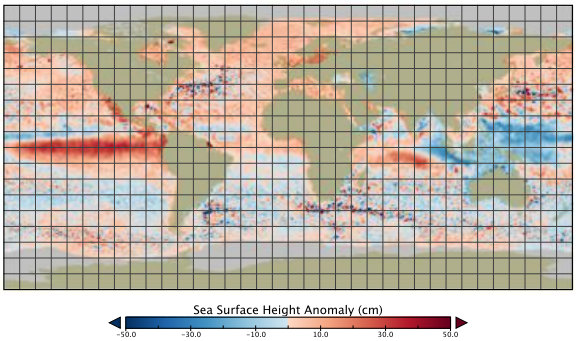

Fig. 11 Global sea surface height (SSH) anomaly for the week of October 10 1997 from the Version 3 dataset of the Archiving, Validation, and Interpretation of Satellite Oceanographic (AVISO) dataset. Eddies can be observed globally as closed contoured negative (dark blue; for cyclonic) or positive (dark red; for anti-cyclonic) anomalies. Best seen in color.

In section 2.2, 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 12) 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

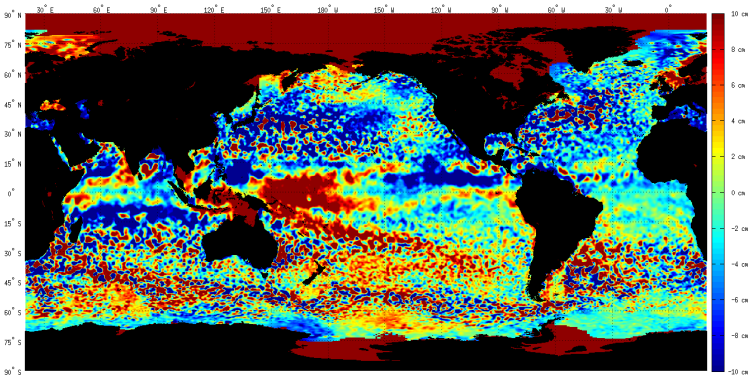

Fig. 12 Global unfiltered SSH anomalies. The data is characterized with high spatial and temporal variability, where values vary widely from one location to the next, as well as across time for the same location. Therefore traditional measures such as mean and standard deviations yield little insight in global patterns.

size unfeasible [40]. 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 [15] making numerous features unnoticeable. False positives are also an issue, as other phenomena such as linear Rossby waves or fronts can masquerade as eddy-like features in SSH data [59, 17]. Finally, given the global and ubiquitous nature of eddies, any learning must be unsupervised. One way to verify the performance of eddy identification and tracking algorithms is to use field-studies data, where floats and ships physical sit on top of eddies. However, such datasets would only provide anecdotal evidence. 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 13 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 favors the largest region possible). Furthermore, such a thresholding approach is known to merge eddies in close proximity [16]. 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 [34]. 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

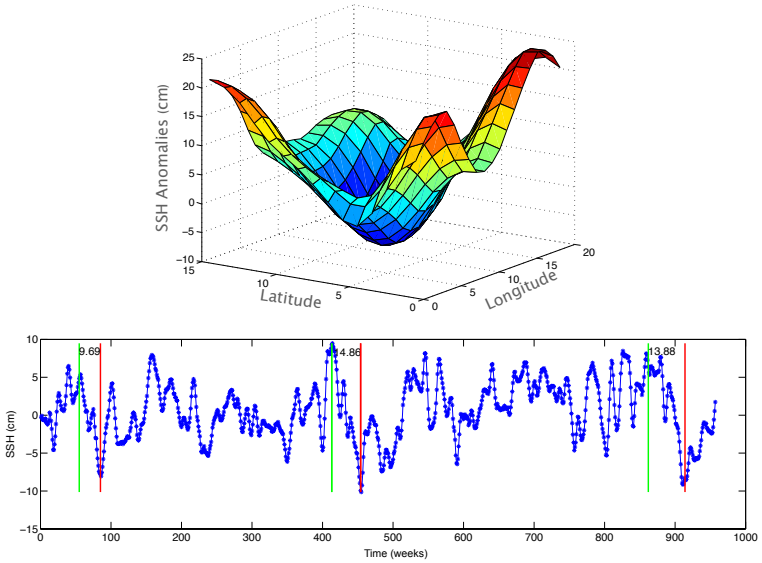

Fig. 13 Two different but complementary views of eddies' effect on SSH anomalies. *Top*: A three dimensional view of a cyclonic eddy in the SSH field. *Bottom*: an SSH time-series at single location. In both cases, the presence of an eddy is indicated through a sustained SSH depression.

and spurious signals. Our method attempts to combine bother approaches to address each method's limitations. We begin by discussing each approach in more detail.

4.1 Spatial methods for ocean eddy identification (threshold-based)

Spatial methods that identify eddies in the SSH field assign binary values to single-time SSH snapshots 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. Given the noise in the SSH field, a second round of pruning occurs after tracking the features across time-frames and discarding any features that did not persist beyond four weeks. Figure 14 shows the ubiquitous cyclonic eddy features identified in a single SSH snapshot. Each snapshot contains a few thousand eddy-like features. However that number is often reduced by a variety of significance tests mentioned earlier.

Chelton *et al.* [15] was the first to track 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 non-linear eddies and linear Rossby waves. In the most comprehensive SSH-based eddy tracking study to date, Chelton *et al.* [16] identified eddies globally as closed contoured smoothed SSH

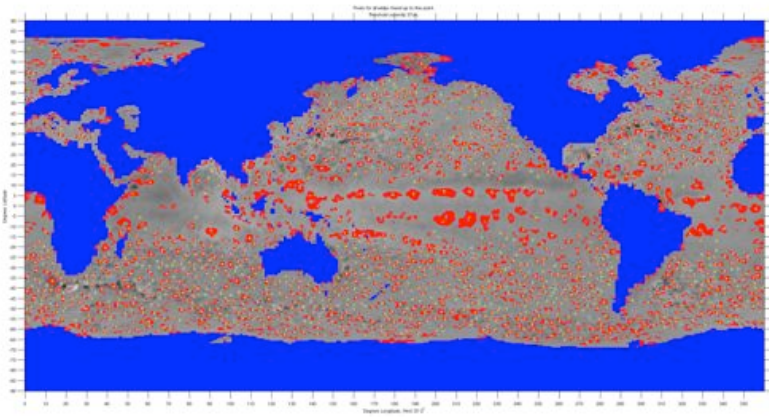

Fig. 14 Eddy-like features are ubiquitous in global SSH data. The challenges is in identifying and tracking such features within a continuous SSH field.

anomalies using a thresholding and nearest neighbor search approach. A similar algorithm was presented in [35] with a few modifications over [16] to improve the runtime complexity and accuracy of the threshold-based method.

At a high level, threshold-based algorithms extract candidate connected components from SSH data by gradually thresholding the data and finding connected component features at each threshold. For each connected component, we applied six criteria to determine if a feature is an eddy candidate: (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; (v) the distance between any two pixels along the contour of the feature must be less than a fixed maximum; and (vi) each connected component must have a predefined convex hull ratio as a function of the latitude of the eddy. The first five conditions are similar to those proposed by [16]. 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 that of Chelton *et al.* [16] are two-fold: First, we use unfiltered data while Chelton *et al.* [16] pre-process the data. Second, to ensure the selection of compact rotating vortices, Chelton *et al.* [16] 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 $O(N)$ in the number of features identified.

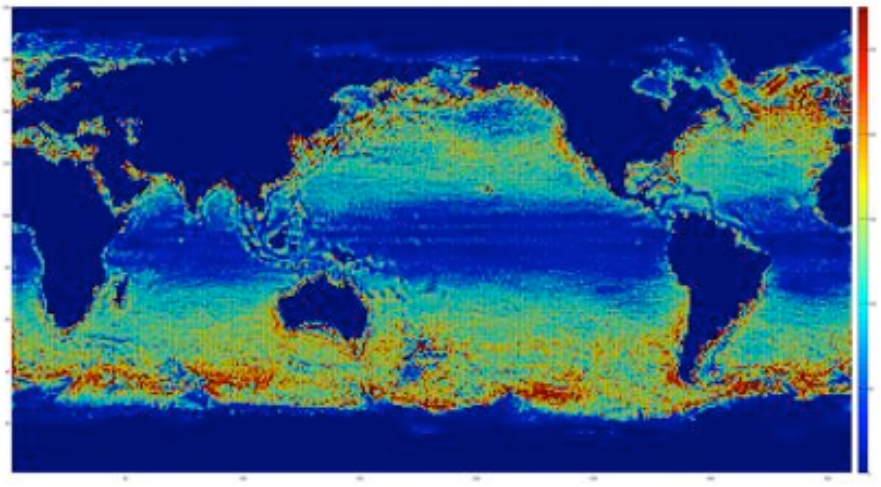

Fig. 15 Aggregate counts for eddy centroids that were observed through each $1^\circ \times 1^\circ$ region over the October 1992 - January 2011 period as detected EddyScan. These results show high eddy activity along the major currents such as the Gulf Stream (North Atlantic) and Kuroshio Current (North Pacific). Best seen in color.

There are instances, however, when the maximum distance criterion is unable to avoid merging several smaller eddies together. Figure 16 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.

4.2 Temporal method for ocean eddy identification

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

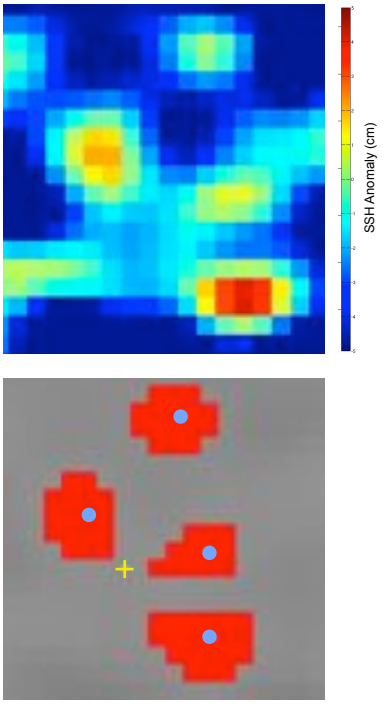

Fig. 16 An example of when Chelton *et al.* [16] maximum distance criterion is met, yet the large feature is in fact several eddies merged together. **Top:** a zoomed-in view on SSH anomalies in the Southern Hemisphere showing at least four coherent structures with positive SSH anomalies. **Bottom:** Chelton *et al.* [16] (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.

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 [12]) 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.

Figure 17 demonstrates how our approach detects candidate time-series. The top panel shows the SSH anomaly time-series for one grid point in the Nordic Sea. For this particular location, our algorithm 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.

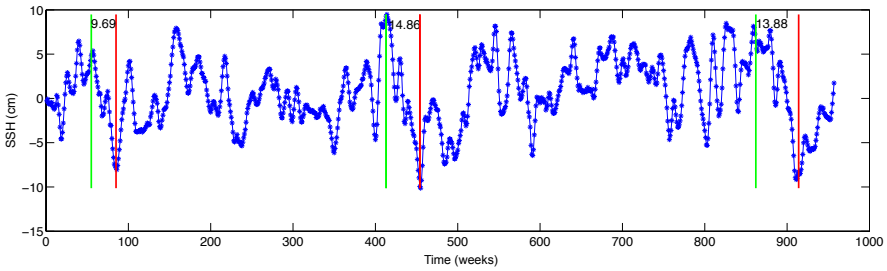

Fig. 17 A sample time-series analyzed by PDELTA with gradually decreasing segments enclosed between each pair of green and red lines. These segments were obtained after discarding segments of very short length or insignificant drop that are atypical signatures of an eddy.

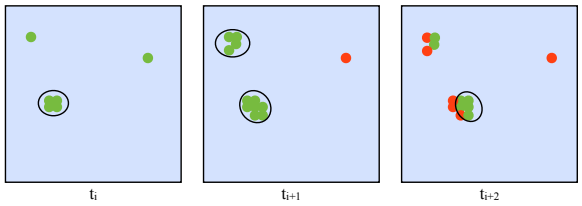

Fig. 18 An illustration to show PDELTA's spatial analysis component. At any given time t_i only a subset of all time-series are labeled as candidates for being part of an eddy (green points). Only when a sufficient number of similarly behaving neighbors are detected (in this case four) PDELTA labels 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).

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 Chelton *et al.* [16]. Overall, we identified a total of 9.08 eddies per month versus 8.87

for Chelton *et al.* [16][5]. This could be due to the fact that eddies tend to be smaller in the region analyzed, and thus could have been ignored by CH11's algorithm once the data were filtered. Figure 19 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.

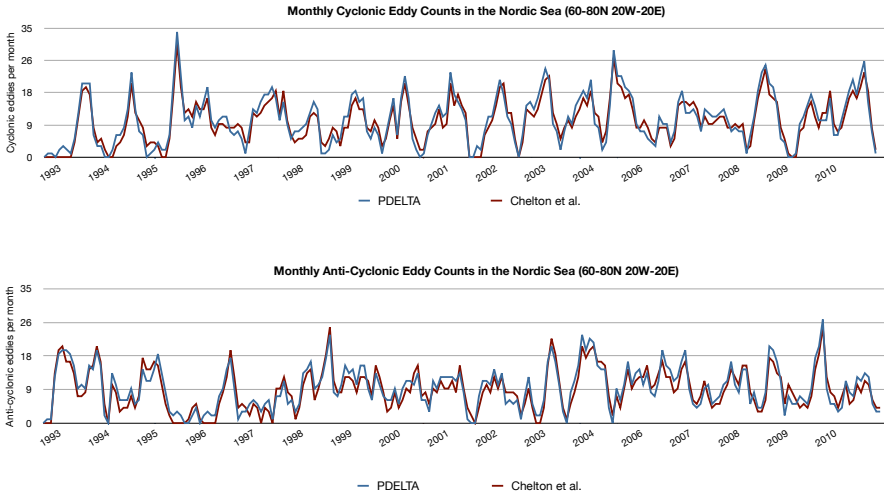

Fig. 19 Monthly eddy counts (lifetime $\geq$ 16 weeks). **Top:** Monthly counts for cyclonic eddies as detected by our automated algorithm PDELTA (blue) and CH11 (red). **Bottom:** Monthly counts for anti-cyclonic eddies as detected by our automated algorithm PDELTA (blue) and Chelton *et al.* [16] (red).

One major advantage of considering the spatio-temporal context of the SSH data is that such an approach scales well with respect to the data's resolution and time-series length (*i.e.* number of satellite snapshots). Figure 20 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.

[5] Data available at: http://cioss.coas.oregonstate.edu/eddies/

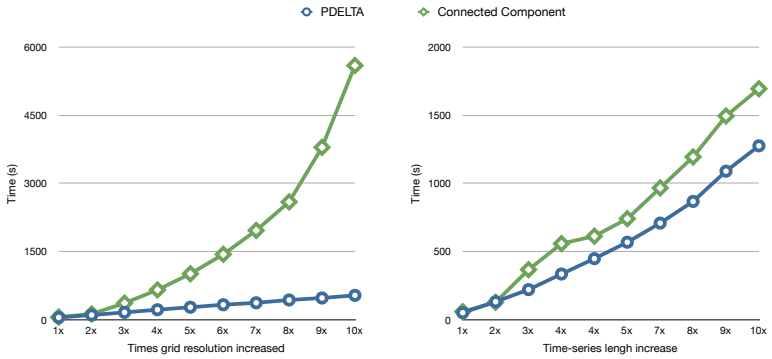

Fig. 20 Scalability comparison between our algorithm PDELTA (blue) and a connected component algorithm (green) similar to CH11. **Left:** time required to track all eddies in the dataset as a function of the grid resolution. **Right:** time required to track all eddies in the dataset as a function of 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.

5 Conclusion and Future Directions

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.

Based on some of the information presented in this chapter, 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. [58]) 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.

Acknowledgements

Part of the research presented in this chapter was funded by an NSF Graduate Research Fellowship, an NSF Nordic Research Opportunity Fellowship, a University of Minnesota Doctoral Dissertation Fellowship, and an NSF Expeditions in Computing Grant (IIS-1029711). Access to computing resources was provided by the University of Minnesota Supercomputing Institute. The authors thank Varun Mithal for generating Figure 4 and Dr. Stefan Sobolowski for generating Figure 7. We also thank Dr. Stefan Liess for constructive comments that improved the quality of the manuscript.

References

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

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

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

[4] Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., *et al.* (2001). 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.

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

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

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

[8] Boriah, S., Kumar, V., Steinbach, M., Potter, C., and Klooster, S. (2008). 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.

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

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

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

[12] Chamber, Y., Garg, A., Mithal, V., Brugere, I., Lau, M., Krishna, V., Boriah, S., Steinbach, M., Kumar, V., Potter, C., and Klooster, S. A. (2011). 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.

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

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

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

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

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

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

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

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

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

[22] Clearwater, S. H. and Provost, F. J. (1990). 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.

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

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

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

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

[27] Domingos, P. (1998). 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[53] Kawale, J., Chatterjee, S., Ormsby, D., Steinhaeuser, K., Liess, S., and Kumar, V. (2012). 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.

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

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

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

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

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

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

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

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

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

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

[64] Mithal, V., Garg, A., Brugere, I., Boriah, S., Kumar, V., Steinbach, M., Potter, C., and Klooster, S. (2011a). Incorporating natural variation into time-series based land cover change identification. In *Proceeding of the 2011 NASA Conference on Intelligent Data Understanding (CIDU)*.

[65] Mithal, V., Garg, A., Boriah, S., Steinbach, M., Kumar, V., Potter, C., Klooster, S., and Castilla-Rubio, J. C. (2011b). Monitoring global forest cover using data

mining. *ACM Transactions on Intelligent Systems and Technology (TIST)*, **2**(4), 36.

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

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

[68] Neill, D. B., Moore, A. W., Sabhnani, M., and Daniel, K. (2005). 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.

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

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

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

[72] Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., and Keogh, E. (2012). 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.

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

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

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

[76] Sencan, H., Chen, Z., Hendrix, W., Pansombut, T., Semazzi, F. H. M., Choudhary, A. N., Kumar, V., Melechko, A. V., and Samatova, N. F. (2011). 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.

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

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

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

[80] Steinbach, M., Tan, P.-N., Kumar, V., Klooster, S., and Potter, C. (2003). 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.

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

[82] Stolorz, P. and Dean, C. (1996). 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.

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

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

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

[86] Team, C. W. (2007). *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.

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

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

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

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

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

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

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

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

[95] Webster, P. J., Holland, G. J., a. Curry, J., and Chang, H. (2005). Changes in tropical cyclone number, duration, and intensity in a warming environment. *Science*, **309**(5742), 1844 –1846.

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

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

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

[99] Wu, E., Liu, W., and Chawla, S. (2008). 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.

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

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

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

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