POTSDAM INSTITUTE
FOR CLIMATE IMPACT RESEARCH

Institute of Physics and Astronomy

Universität Potsdam

Characterizing the spatio-temporal patterns of Extreme Events

from recurrence to prediction

Abhirup Banerjee

May, 2022

Cumulative dissertation
to obtain the academic degree
"doctor rerum naturalium" (Dr. rer. nat.)
in the scientific discipline of Theoretical physics
Date of disputation 16.08.2022

Submitted to the
Faculty of Mathematics and Natural Sciences
at the University of Potsdam, Germany

Supervisors
Prof. Dr. Jürgen Kurths
PD Dr. Norbert Marwan
Prof. Dr. Bruno Merz

Reviewers
Prof. Dr. Jürgen Kurths
PD Dr. Norbert Marwan
Prof. Pinaki Pal

Examination Committee Members
Prof. Dr. Jürgen Kurths
PD Dr. Norbert Marwan
Prof. Pinaki Pal
Prof. Dr. Bruno Merz
Prof. Dr. Karoline Wiesner
Prof. Dr. Carsten Beta

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-55983
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-559839

To Megh, Godai, Sweety and Rio

"*As you start to walk on the way, the way appears.*"
– Rumi

Acknowledgements

My path in research has been a long, arduous journey filled with excitement and disappointments alike. However, I am fortunate to have encountered some incredible people along the way who have deeply motivated me and lent me their support at every step I have taken ahead. The lessons I learned from each of them molded me into what I am today.

First and foremost, I am deeply indebted to my supervisor, Prof. Jürgen Kurths, for being that strong and big support under whose shade I confidently grew as a researcher. I would also like to express my utmost gratitude to my day-to-day supervisor, Dr. Norbert Marwan, who gave me ample freedom to explore my curiosities while also enduring my naive questions. He showed me a way to develop critical scientific thinking and inspired me to continue further research work. In him, I found a wonderful companion for discussing my scientific ideas, in particular, philosophy of science in general. I am also highly grateful to my co-supervisor, Prof. Bruno Merz, for his constructive criticism and insightful comments on my work which motivated me to discover the utilities of my methodology in hydrology and climate science. I am very thankful to all my supervisors for giving me this opportunity which has opened new avenues in my life. But, this would not have been possible without two very important people, my uncle Prof. Samit Chattopadhyay, and my teacher Dr. Syamal Dana. I would like to express my heartfelt gratitude to them for motivating me to pursue scientific research. Syamal Sir has been the guiding star of my research career path whose advice I deeply rely on.

I acknowledge the support from Deutsche Forschungsgemeinschaft (DFG) under the NatRiskChange research training group for my Ph.D. I truly appreciate the knowledge I gained from all the members of the NatRiskChange group through the formal and informal scientific discussions during our collaborations, group projects, and retreats.

Many thanks to the Potsdam Institute for Climate Impact Research (PIK) for providing such a great environment for research and interaction with a plethora of highly motivated researchers. I would like to thank Till Hollmann, Sophia Kostial, Gabriele Pilz, and the welcome center for their enormous help with the highly complex administrative work. I owe a special thanks to Dr. Bedartha Goswami for his extensive help, support, and guidance during my early days of research work. I express my heartfelt thanks to Shraddha Gupta who motivated me towards the field of climate networks and to grow a deep interest in understanding the climate processes. She has always been critically appreciative of my works and provided a lot of support and valuable inputs. I would like to thank Dr. Yang Liu and Zhen Su, as my research buddies for their help regarding programming and computation. A special thanks to my lab partner and friend, Tobias Braun, for engaging in various scientific and non-scientific discussions. I am also thankful to Prof. Ankit Agarwal for his encouragement and support as a senior colleague. I am extremely grateful to my colleague, friend, and neighbor, Shubham Sharma, for his company during the pandemic.

I would like to thank my colleagues, Lukas Schoppa and Jonas Wassmer, for helping me with German translations. I owe a special thanks to Chittaranjan Hens, Sadhitro De, and Ravi Kumar Guntu for proofreading my thesis and providing valuable remarks.

I would like to thank my family for their emotional support that propelled me to continue my journey. I am extremely grateful to all my teachers who helped me to develop a creative curiosity. Lastly, I am thankful to myself, for persevering through the difficult times of the pandemic and finally reaching this day.

Abstract

Over the past decades, there has been a growing interest in 'extreme events' owing to the increasing threats that climate-related extremes such as floods, heatwaves, droughts, etc., pose to society. While extreme events have diverse definitions across various disciplines, ranging from earth science to neuroscience, they are characterized mainly as dynamic occurrences within a limited time frame that impedes the normal functioning of a system. Although extreme events are rare in occurrence, it has been found in various hydro-meteorological and physiological time series (e.g., river flows, temperatures, heartbeat intervals) that they may exhibit recurrent behavior, i.e., do not end the lifetime of the system. The aim of this thesis to develop some sophisticated methods to study various properties of extreme events.

One of the main challenges in analyzing such extreme event-like time series is that they have large temporal gaps due to the paucity of the number of observations of extreme events. As a result, existing time series analysis tools are usually not helpful to decode the underlying information. I use the *edit distance* (ED) method to analyze extreme event-like time series in their unaltered form. ED is a specific distance metric, mainly designed to measure the similarity/dissimilarity between point process-like data. I combine ED with recurrence plot techniques to identify the recurrence property of flood events in the Mississippi River in the United States. I also use recurrence quantification analysis to show the deterministic properties and serial dependency in flood events.

After that, I use this non-linear similarity measure (ED) to compute the pairwise dependency in extreme precipitation event series. I incorporate the similarity measure within the framework of complex network theory to study the collective behavior of climate extremes. Under this architecture, the nodes are defined by the spatial grid points of the given spatio-temporal climate dataset. Each node is associated with a time series corresponding to the temporal evolution of the climate observation at that grid point. Finally, the network links are functions of the pairwise statistical interdependence between the nodes. Various network measures, such as *degree*, *betweenness centrality, clustering coefficient,* etc., can be used to quantify the network's topology. We apply the methodology mentioned above to study the spatio-temporal coherence pattern of extreme rainfall events in the United States and the Ganga River basin, which reveals its relation to various climate processes and the orography of the region.

The identification of precursors associated with the occurrence of extreme events in the near future is extremely important to prepare the masses for an upcoming disaster and mitigate the potential risks associated with such events. Under this motivation, I propose an in-data prediction recipe for predicting the data structures that typically occur prior to extreme events using the Echo state network, a type of Recurrent Neural Network which is a part of the reservoir computing framework. However, unlike previous works that identify precursory structures in the same variable in which extreme events are manifested (*active* variable), I try to predict these structures by using data from another dynamic variable (*passive* variable) which does not show large excursions from the nominal condition but carries imprints of these extreme events. Furthermore, my results demonstrate that the quality of prediction depends on the magnitude of events, i.e., the higher the magnitude of the extreme, the better is its predictability skill. I show quantitatively that this is because the input signals collectively form a more coherent pattern for an extreme event of higher magnitude, which enhances the efficiency of the machine to predict the forthcoming extreme events.

Keywords – Extreme events, Floods, Extreme precipitation, Edit distance, Recurrence plot, Complex network, Echo state network.

Zusammenfassung

In den letzten Jahrzehnten hat das Interesse an 'Extremereignissen' aufgrund der zunehmenden Bedrohung, die klimabedingte Extreme wie Überschwemmungen, Hitzewellen, Dürren usw. für die Gesellschaft darstellen, zugenommen. Obwohl Extremereignisse in verschiedenen Disziplinen - von der Geowissenschaft bis zu den Neurowissenschaften - unterschiedlich definiert werden, werden sie hauptsächlich als dynamische Ereignisse innerhalb eines begrenzten Zeitrahmens charakterisiert, die das normale Funktionieren eines Systems beeinträchtigen. Obwohl Extremereignisse selten vorkommen, wurde festgestellt, dass verschiedene hydro-meteorologische und physiologische Zeitreihen (z. B. Stromabflussmengen, Temperaturen, Herzschlagintervalle) rekurrentes Verhalten. Das heißt, sie enden nicht an der Lebensdauer des Systems. Das Ziel dieser Arbeit ist es, fortschrittliche Methoden zur Untersuchung verschiedener Eigenschaften von Extremereignissen zu entwickeln.

Eine der größten Herausforderungen bei der Analyse solcher extremen Ereignisse ist, dass sie große zeitliche Lücken aufweisen, da die Anzahl beobachteter Extremereignissen gering ist. Bestehende Zeitreihenanalysetools sind daher in der Regel nicht hilfreich, um die zugrundeliegenden Informationen zu entschlüsseln. Ich verwende die Edit-Distanz (ED) Methode, um extremeereignisähnliche Zeitreihen in ihrer unveränderten Form zu analysieren. ED ist eine spezielle Abstandsmetrik, die hauptsächlich zur Messung der Ähnlichkeit/Unähnlichkeit zwischen punktprozessähnlichen Daten entwickelt wurde. Ich kombiniere ED mit Rekurrenzplots, um die Wiederkehr-Eigenschaften von Hochwasserereignissen am Mississippi in den Vereinigten Staaten zu ermitteln. Außerdem werte ich die Wiederkehr-Eigenschaften quantitativ aus, um die deterministische Charakteristik und die serielle Abhängigkeit bei Hochwasserereignissen aufzuzeigen.

Anschließend verwende ich dieses nichtlineare Ähnlichkeitsmaß(ED), um die paarweise Abhängigkeit in extremen Niederschlagsereignisreihen zu berechnen. Ich verknüpfe das Ähnlichkeitsmaß mit der Theorie komplexer Netzwerke, um das kollektive Verhalten von Klimaextremen zu untersuchen. In diesem Fall werden die Knoten durch die räumlichen Gitterpunkte des zu untersuchenden raumzeitlichen Klimadatensatzes definiert. Jeder Knoten repräsentiert eine Zeitreihe, die die zeitliche Entwicklung der Klimabeobachtung an diesem Gitterpunkt beschreibt. Schließlich sind die Netzwerkverbindungen Funktionen der paarweisen statistischen Interdependenz zwischen den Knotenpunkten. Zur Quantifizierung der Netztopologie können verschiedene Netzwerkmaße wie unter anderem der Grad, die Betweenness-Zentralität, oder der Clustering-Koeffizient verwendet werden. Wir wenden die oben erwähnte Methodik an, um das raumzeitliche Kohärenzmuster extremer Niederschlagsereignisse in den Vereinigten Staaten und im Einzugsgebiet des Ganges zu untersuchen. Die Ergebnisse zeigen, dass ein Zusammenhang zwischen dem Kohärenzmuster und verschiedenen Klimaprozessen und der Orographie der Region besteht.

Die Identifizierung von Vorläufern, die mit dem Auftreten von Extremereignissen in naher Zukunft verbunden sind, ist äußerst wichtig, um die Bevölkerung auf eine bevorstehende Katastrophe vorzubereiten und potenziell resultierende Risiken zu mindern. Deshalb schlage ich ein datenbasiertes Vorhersageverfahren zur Bestimmung der Datenstrukturen vor, die typischerweise vor extremen Ereignissen auftreten. Das Verfahren basiert auf dem Echo-State Netzwerk, einem rekurrenten neuronalen Netz, das dem Reservoir-Computing zugeordnet wird. Im Gegensatz zu früheren Arbeiten, die Vorläuferstrukturen in der Variablen identifizieren, in der

sich Extremereignisse manifestieren (aktive Variable), versuche ich die Strukturen anhand anderer dynamischer Variablen (passive Variablen) vorherzusagen. Diese Variablen weichen selbst nicht sonderlich von ihrem eigenen Sollzustand ab, aber sie besitzen eine Aussagekraft gegenüber den Extremereignissen. Meine Ergebnisse zeigen, dass die Qualität der Vorhersage von der Magnitude der Ereignisse abhängt, d.h. je extremer ein Ereignis ist, desto besser kann es vorhergesagt werden. Ich belege quantitativ, dass dieser Zusammenhang darauf basiert, dass die gesammelten Eingangssignale aussagekräftigere Kohärenzmuster für Extremereignisse hoher Magnitude bilden. Dies erhöht die Wirksamkeit des Ansatzes bei der Vorhersage bevorstehender Extremereignisse.

Schlüsselwörter – Extremereignisse, Überschwemmungen, Extremniederschläge, Edit-Distanz, Rekurrenzplot, komplexes Netzwerk, Echo-State Netzwerk.

Contents

List of publications

This dissertation is based on the following publications.

P1. **Banerjee, A.**, Goswami, B., Hirata, Y., Eroglu, D., Merz, B., Kurths, J., and Marwan, N.: Recurrence analysis of extreme event-like data, Nonlin. Processes Geophys., 28, 213-229, https://doi.org/10.5194/npg-28-213-2021, 2021.

P2. Agarwal, A., Guntu, R K., **Banerjee, A.**, Gadhawe, M A., and Marwan, N., A complex network approach to study the extreme precipitation patterns in a river basin, Chaos 32, 013113 (2022) https://doi.org/10.1063/5.0072520

P3. **Banerjee, A.**, Kemter, M. , Goswami, B., Merz, B., Kurths, J., Marwan, N: Spatial coherence patterns of extreme winter precipitation in the United States, Under review at *Theoretical and Applied Climatology*, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-1243541/v1]

P4. **Banerjee, A.**, Mishra, A. , Dana, S K. Hens, C., Kapitaniak, T., Kurths, J., Marwan, N: Predicting the data structure prior to extreme events from passive observables using Echo State network, Submitted to *Frontiers in Applied Mathematics and Statistics*

Further collaboration work (not included in this dissertation)

P5. Kemter, M., Fischer, M., Luna, L. V., Schönfeldt, E., Vogel, J., **Banerjee, A.,** et al. Cascading hazards in the aftermath of Australia's 2019/2020 Black Summer wildfires, Earth's Future

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

The term 'extreme event' (EE) is very commonly known and used across various scientific/non-scientific disciplines, particularly, in recent years due to the occurrence of extreme weather events. Among the diverse definitions associated with the term, the events which have been labeled as EEs in the past include droughts (Ebi and Bowen, 2016), tsunamis (Annex, 2012), cyclones, earthquakes, acts of terrorism (Comfort, 2002), epidemics, power outages (Jentsch et al., 2006), epileptic seizures (Nadin, 2006), etc. The definition of an extreme event is not well established as it depends on the disciplines in which the term is used (McPhillips et al., 2018; Broska et al., 2020; Jentsch et al., 2006). The study of extreme events has gained attention due to the disastrous and widespread impacts of such events in various fields like sociology (Webb, 2002), ecology (Gutschick and BassiriRad, 2003), finance (Embrechts et al., 2013), share markets (Krause et al., 2015a), weather and climate (Sura, 2011), to name but a few. Climate extremes such as high temperature (heat waves), extreme precipitation, drought, or flooding are generally defined as the reaching a level or value above (or below) a threshold (Seneviratne et al., 2012). The frequency and magnitude of climate extremes have increased due to climate change (IPCC, 2018; Kharin et al., 2018; Fischer et al., 2021; McPhillips et al., 2018). Such EEs have a significant impact on our society; for example, Hurricane Harvey caused devastating floods leading to over 80 fatalities and huge economic losses (Van Oldenborgh et al., 2017), the 2021 flood in Western Europe killing around 220 people (Else, 2021), and the mega-heatwaves of 2003 and 2010 in Europe causing widespread crop loss, wildfires, air pollution, transport disruptions and water scarcity (Miralles et al., 2014).

Two of the characteristics of EEs are rarity and unpredictability (Jentsch et al., 2006) – these large events appear from nowhere, without any indication (Farazmand and Sapsis, 2019). However, in some cases, it has been possible to detect early warning signals (Boers et al., 2014a; Carrara et al., 1999; Scheffer et al., 2009; Denny et al., 2009) of such events. A suitable framework is required to study and understand the underlying mechanism of such catastrophic events. Statistical analysis of EEs (Coles, 2001; De Haan et al., 2006) has been shown to be a very useful way to understand the probabilities associated with extreme events (Ghil et al., 2011; Katz et al., 2002; Smith, 2003; Tabari, 2021; Majumdar et al., 2020). Though the appearance of EEs may seem random, but in reality, they are the manifestation of a high dimensional complex system and carry some intricate hidden patterns (Jentsch et al., 2006; Altmann and Kantz, 2005). The goal of this thesis is to unravel the recurrent behavior of these extreme events, understand their spatio-temporal coherence patterns, if any exist, and finally use this knowledge to formulate a recipe to predict EEs.

In the next section, I briefly explain the main aspects of the theoretical foundation of this thesis. The subsequent chapters build upon each of these theoretical aspects, as indicated later, and apply the knowledge to investigate real world phenomena.

1.2 Foundation of the thesis

1.2.1 Handling extreme event-like data

The occurrence of EEs is highly irregular, and often there are large temporal gaps (zeros between two events) present in between the events (Fig.1.1). Analyzing such EE-like time

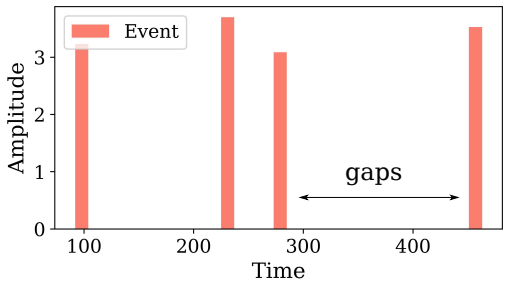


Figure 1.1: Schematic of an event series, bars indicate the events.

series using standard linear/non-linear time series analysis methods is problematic. To handle this problem, I use a method called the *edit distance* (ED), which is, particularly, designed to analyze point process-like data. This method was first introduced by Victor and Purpura (1997) to compute similarity/dissimilarity in spike trains. Later, Hirata and Aihara (2009) extended this method and used it as a metric to represent a spike train as a real-valued time series and coined the term '*edit distance*' to refer to this method. ED is a specific distance metric which measures the similarity between spike trains by transforming one spike train to another and by using some elementary operations, such as deletion of a spike, insertion of a spike, shifting of a spike in time, and adjusting the event amplitude. A certain cost is assigned to each of these operations. The transformation is done in such a way so as to minimize the total cost. Naturally, a lower transformation cost implies higher similarity between the spike trains and vice versa. Therefore, the transformation cost is a measure of the degree of similarity. By incorporating ED with recurrence plot analysis, Suzuki et al. (2010) studied the recurrence property in exchange tick data. Thereafter, Ozken et al. (2015, 2018) used a more compact version of ED to study the recurrence behavior in paleoclimate data.

As mentioned earlier, by definition, EEs are scarce, i.e., there are big gaps between two data points. A very elegant formulation of characterizing extreme value-like behavior, such as that observed in daily rainfall time series, can be derived from the theory of point processes. Therefore, ED gives a good starting point to analyze such data in its unaltered form. In my study, I used ED as a similarity metric to characterize the temporal pattern of EE-like data, such as that of flood events in the Mississippi River in the USA, obtained from river discharge data (Chapter. 2), and spatio-temporal pattern of extreme precipitation time series (Chapter. 4) in the USA during the winter season. The detailed mathematical description of ED is described in Chapter. 2. In the following subsections, I briefly introduce the frameworks that I have used in combination with ED to study the temporal and spatial coherent patterns of EEs.

1.2.2 Temporal patterns of extreme events

Recurrence of events is ubiquitous in nature – we find examples of recurrences from the Earth system to physiological systems (Marwan, 2003). The human brain also has a strong sense of recurrence, ranging from the feeling of *déjá vu*, related to the recollection of a similar experience or feeling, to that associated with neurological illnesses such as epilepsy and hallucinations. French mathematician Henri Poincaré (1854-1912) formulated this phenomenon in his *recurrence theorem* (Poincaré, 1890), which laid the foundation for the further development of this

idea. Eckmann et al. (1987) showed the recurrence of states in a dynamical system using a

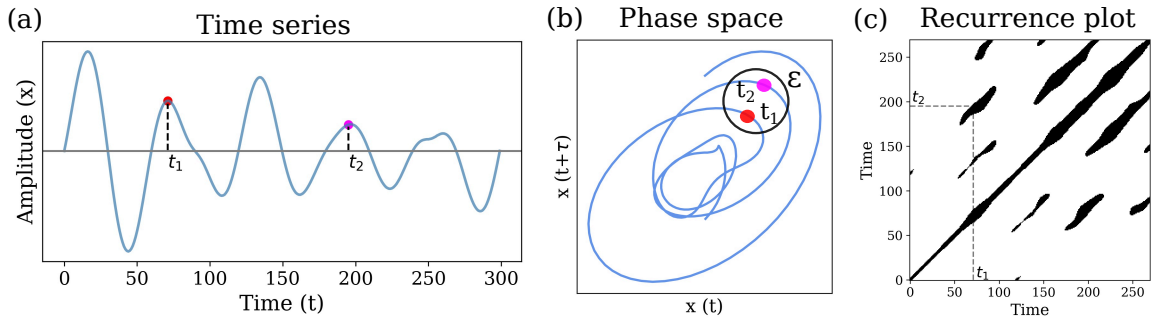


Figure 1.2: Schematic representation of computation of recurrence plot (RP) for a scalar time series. (a) Scalar time series, t_1 and t_2 are two time points. (b) Reconstructed phase space of the time series, the state vectors are constructed using delay embedding. (c) RP of the phase space trajectories – when the two states at time t_1 and t_2 come arbitrarily close, under certain threshold ε, we assign a black dot in the RP.

two-dimensional plot, called 'Recurrence plot' (RP), which encoded the visual topology of the dynamics. RP is a visual tool that graphically shows the times at which the phase space trajectory revisits roughly the same area in phase space (Fig.1.2). The visual appearance of an RP can provide a brief idea about the underlying dynamics of the system. Black dots in the RP represents the recurrence of a state, and the white dots mean no recurrence. The formation of different kinds of structures, such as diagonal lines, vertical and horizontal lines, box structures, etc., in the RP, are caused by the characteristic behavior of the phase space trajectory. The *diagonal lines* are some of the most common line structures (Fig.1.2c). Apart from the main diagonal line (Fig.1.2c), which shows the recurrence of a state with itself, we also find some short diagonal lines in the RP.

The diagonal lines are formed when there are periods in which trajectories evolve arbitrarily close to each other. One can quantify different kinds of line structures to extract specific information about the system (Marwan et al., 2007). To explain simply, in Fig.1.3, I take three different kinds of signals: a periodic signal (Fig.1.3a), a chaotic signal (Fig.1.3b), and a random noise signal (Fig.1.3c). The RP of the periodic signal (Fig.1.3d) contains long, uninterrupted, diagonal lines. The diagonal lines in the RP represent the *deterministic* nature of a system (Marwan et al., 2007). As the periodic signal is highly deterministic, we get well defined diagonal line structures in the RP. Chaotic systems exhibit quasi periodicity, and therefore, the RP of the chaotic signal contains short, interrupted diagonal lines (Fig.1.3e). In the case of a random signal, all the states have an equal probability of occurrence, hence the points (black dots) are uniformly distributed in the RP (Fig.1.3f). Due to its easy computation technique and powerful ability to characterize the underlying dynamics of a system, RP has become a popular tool in the *nonlinear time series* (Kantz and Schreiber, 1997) analysis. RP has been used in a wide range of disciplines, such as Neuroscience, Economics, Hydrology, Geoscience, to name a few (Marwan, 2008) (concise list of articles can be found here [1]).

Considering all these powerful aspects of RP, we can ask the following question: Can RP be used to characterise the repetitive nature of EEs? The answer is yes but not directly. One needs to use some tricks before applying RP to EEs. In the standard definition of RP, we compute the similarity between states by measuring the distance between them by using certain norms, e.g., the Euclidean norm. In the Euclidean norm, the distance between states is computed point-wise. However, in the case of EE-like time series, which have a lot of zeros between two events, the standard method of computing RP would increase the chances of false recurrence. To circumvent this problem, I incorporate ED (Sec.1.2.1) with the RP approach. By implementing

[1] http://www.recurrence-plot.tk/bibliography.php

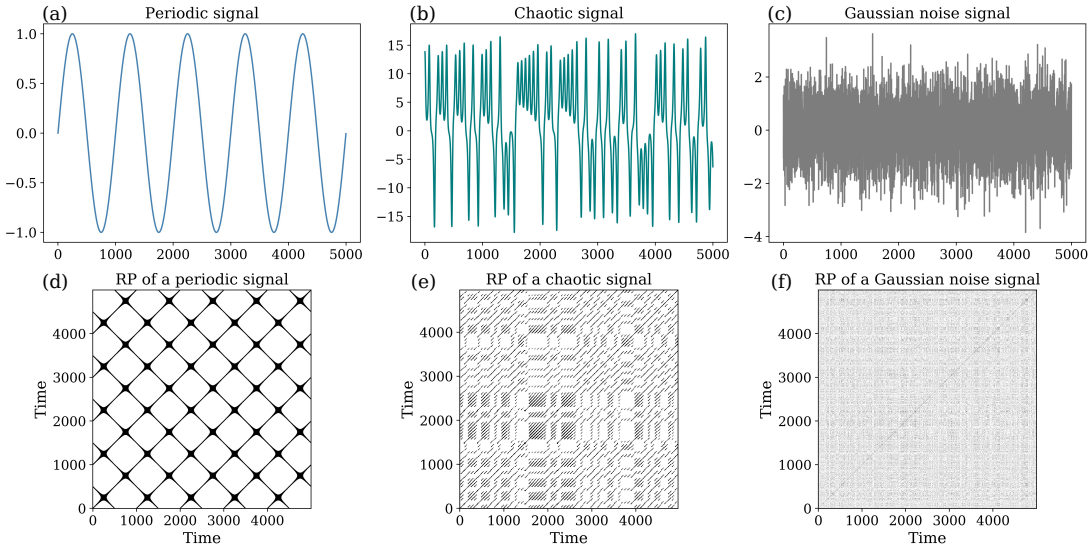


Figure 1.3: Different dynamics, different RP; (a) periodic signal, (b) chaotic signal, (c) Gaussian noise signal, and (b), (e), and (f) are the corresponding RP.

this technique in Chapter. 2, I compute the RP of flood events in the Mississippi river, USA. Using suitable quantification methods, I am able to show that flood events are deterministic in nature and show a long-term correlation.

1.2.3 Spatial coherence patterns of extreme events

The term 'network' refers to a large system consisting of many similar parts that are connected together to allow movement or communication between or along the parts, for example, *social networks* such as Facebook, where we stay connected with our friends, family, colleagues via the internet. There is a plethora of examples of such 'networks' in various disciplines (Newman, 2010). Complex network theory is a branch of science which studies the relationships between different components of a complex system. Complex network theory has become an agile tool to study the emergent properties of complex systems in various areas like physics, biology, economics, social sciences, to name a few (Watts and Strogatz, 1998; Newman, 2010; Barrat et al., 2008; Barabási, 2013). The foundation of complex networks is based on the graph theory. A graph G consists of two elements, a set of nodes or vertices v, and a set of links/ edges e connecting the nodes; $G = \{v, e\}$ (Fig. 1.4). During the last decade, the application of

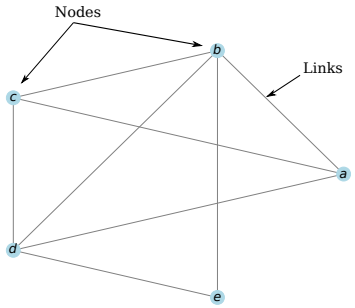


Figure 1.4: A simple graph containing a collection of nodes (a),(b),...,(e) that are connected by links (black lines).

complex network theory to climate science, referred to as *climate network*, has been extensively used to study various climate phenomena such as the El Niño-Southern Oscillation and the monsoon (Tsonis and Roebber, 2004; Tsonis et al., 2006; Fan et al., 2021b).

Climate network belongs to a class of *functional networks* (Donges et al., 2011; Dijkstra et al., 2019). There is a fundamental difference between *structural networks*, e.g., power grids, internet,

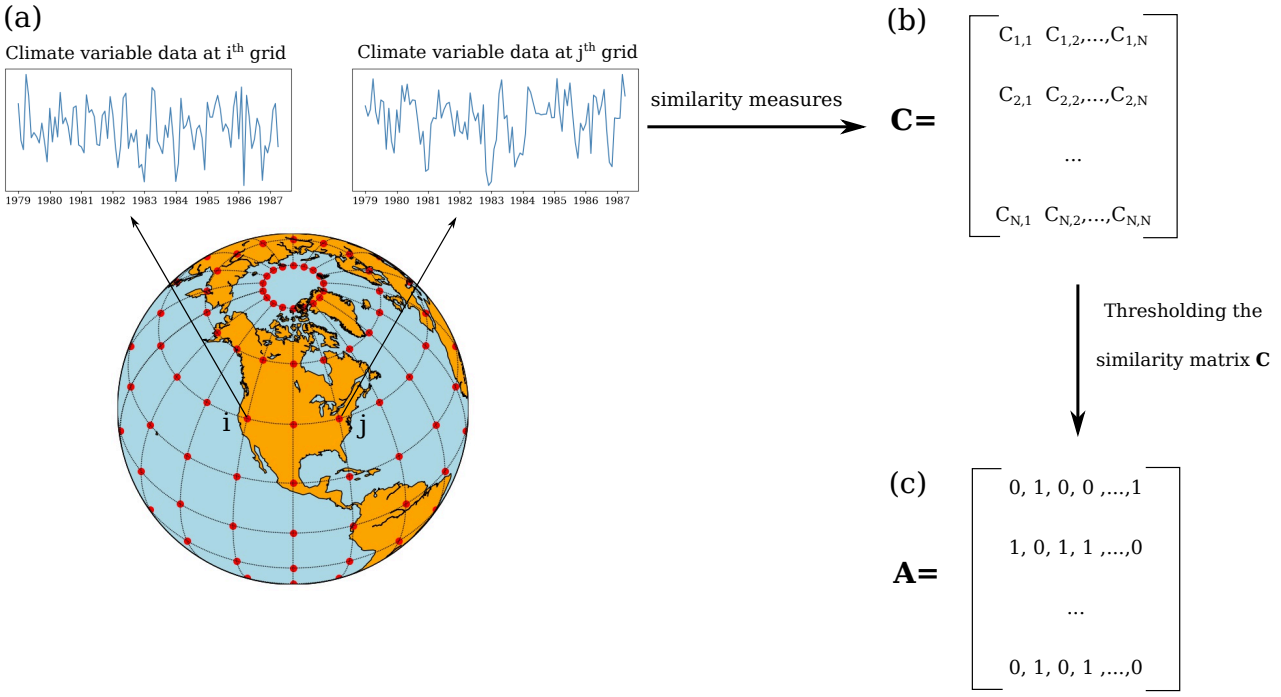


Figure 1.5: Schematic representation of climate network construction.(a) Geographical grid points (red dots on the globe), consisting of climate variable time series, are considered as nodes, (b) by using a suitable similarity measure (Fan et al., 2021b), we construct the similarity matrix **C** by computing pairwise interdependence, (c) adjacency matrix **A** is obtained by thresholding the **C**.

and *functional networks*, e.g., climate network, and brain network (Zhou et al., 2006). In the former, the links are physically evident connections between the nodes, whereas, in the latter, links represent associations or functional dependencies between the nodes. The construction of the (functional) climate network is shown in Fig.1.5. Let's assume we have a spatiotemporal dataset $\mathbf{Z} \in \mathbb{R}^{N \times T}$, where each node $i \in N$ is associated with a time series that encodes the temporal evolution (T_i) of the climate variable (Fig.1.5a). The geographical grid points of the gridded spatiotemporal climate dataset are identified as the nodes of the network (here i and j in Fig.1.5a). The statistical dependency between every pair of nodes, i and j, is computed using an appropriate similarity measure. This gives the similarity matrix C_{ij} (Fig.1.5b). A suitable threshold is used to detect the strongest or most statistically significant dependencies to construct the adjacency matrix **A** (Fig.1.5c) which encodes the interactions or links in the network. The network topology is represented by the adjacency matrix, where $\mathbf{A}_{ij} = 1$ (for unweighted and undirected network) indicates the connection between two nodes i and j, and $\mathbf{A}_{ij} = 0$ denotes no connection between the nodes. Thereafter, one can use different network measures to characterize the topology of the network.

One of the key features of the climate network approach is that, it provides a comprehensive picture of the underlying dynamics. For example, it can capture the signatures of large-scale modes of climate variability, such as the ENSO (Yamasaki et al., 2008; Gozolchiani et al., 2011) or atmospheric teleconnections (Boers et al., 2019). It has shown a great potential to investigate the spatially synchronous patterns of extreme events, such as extreme precipitation (Boers et al., 2013, 2014c; Malik et al., 2011), heat waves (Mondal and Mishra, 2021). Understanding the spatial extent of these climate extremes and the impacts of global climate change has become a serious concern as these events can trigger other natural hazards. Spatially correlated hazards on a regional scale, including heavy precipitation and wind extremes, can also be caused by

individual weather systems like atmospheric blockings or storms - such as Lothar (Wernli et al., 2002) and Ophelia (Guisado-Pintado and Jackson, 2018) in Europe. Extreme synchronous precipitation could lead to river floods (Merz et al., 2021), landslides (Gariano and Guzzetti, 2016), etc.

In Chapter. 3 and Chapter. 4, I study the spatio-temporal coherence patterns of extreme precipitation in India, and in the United States during winters, using ED as a similarity measure to construct the extreme precipitation's network (Fig.1.5). By using suitable network measures, I try to analyze network topology to identify spatially coherent regions of extreme precipitation. I also try to find the connection between the network patterns with different climate processes and the topography of the region. I also compare ED with other similarity measure such as the event synchronization (Malik et al., 2012) in Chapter. 3. The application of ED as a similarity measure in the climate network framework is advantageous over other measures as the method can successfully incorporate the strength of the EE along with their time of occurrence while measuring the similarity, unlike the previous works (Malik et al., 2012; Boers et al., 2013; Wolf et al., 2020).

1.2.4 Prediction scheme of extreme events

The prediction of extreme events has earned significant attention in various disciplines due to the damaging impact associated with these events. But unfortunately, not much progress has been made in developing a robust framework to predict such events because of the complexity of the associated dynamics (Farazmand and Sapsis, 2019). For centuries, weather prediction has been of particular interest due to its immense relevance in all aspects of daily life, such as agriculture, transportation, etc. In recent times, high performance computing resources and more sophisticated methods have been developed to improve weather forecasts. But we are still unable to predict future extremes, e.g., the Great Storm of October 1986 in South England, and the Elbe river flood (2002) in Germany caused by extreme precipitation, due to the rarity of occurrence of such events and the high complexity of the Earth's climate system. Prediction of EEs also requires detailed knowledge of the present state of the system, which is usually unavailable. The partial knowledge of the current state together with the chaotic nature of the system leads to uncertainty in the future predictions (Murphy et al., 2004; Mohamad and Sapsis, 2015).

We can avoid or prepare ourselves from the undesirable impacts caused by EEs if we can get an early warning before the occurrence of EEs. Anticipating forthcoming events based on some precursors is a very old idea. In many cultures across the world, superstitions such as a black cat crossing your path or the breaking of glass are regarded as bad omens while sighting a pair of mynahs is supposed to be a harbinger of good luck. However, in case of superstitions, these precursors may have false correlations with extreme events. On the other hand, there are effective precursors such as unusual animal behaviour like that of birds before a storm or mice before a plague, and the old adage "*Red sky at night, sailor's delight. Red sky in the morning, sailor's warning*," have been passed on as common knowledge. There have been many attempts to explain these precursory signals scientifically, for instance studies show that some animals are more capable of perceiving geographic stimuli such as seismic or acoustic waves at low frequency (below 50 Hz), electric field changes, etc., which possibly can explain their anomalous behaviour prior to earthquakes (Buskirk et al., 1981) or hurricanes (Heckscher, 2018). The search for precursory signals to predict extreme events is an extremely important topic of research in various fields such as seismology, for the prediction of earthquakes (Ouzounov et al., 2018a), and in neuroscience, for the prediction of epileptic seizures (Rings et al., 2019a). It is interesting to note that often these precursory signals are observed in some other subsidiary variables or components of the system instead of the ones in which extreme events are visibly

observed. For instance, precursory information about earthquake events can be obtained from electromagnetic signals (Karamanos et al., 2006), and epileptic seizures could be forecasted using heart rate signals whose high prediction sensitivity matches with forecasting using EEG signals (Kerem and Geva, 2005). Furthermore, studies using complex network-based methods have previously shown that seizure precursors are entirely associated with brain regions that are deemed unaffected by the focal epileptic process (Rings et al., 2019b). This is because, EEs observed in a variable are a manifestation of the rich dynamics of a multivariate higher dimensional complex system. As an example, a flood is an extreme event observed in discharge data, but the occurrence of a flood depends on other variables, such as precipitation, soil moisture, temperature, etc. Though the subsidiary variables may have a slower variability and do not exhibit prominent extreme behaviour (*passive*), they carry significant information about them. This points out to the importance of prediction of EEs through a systematic identification of precursory patterns, using data driven methods, in not only those variables which are visibly affected by EEs (*active*) but also in other *passive* variables.

Most previous works have suggested data-driven methods to predict a pattern in the data structure, especially, if it exists as a coherent pattern, preceding an extreme event in the active variable using the variable itself as input. However, the discussion in the aforementioned paragraph motivated me to build a framework to predict the data structure or pattern in the active variable prior to the appearance of an extreme event from passive variables as input data source using machine learning methods.

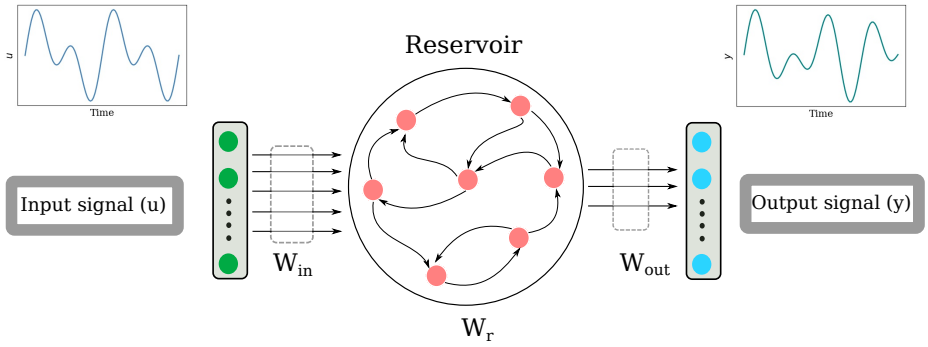


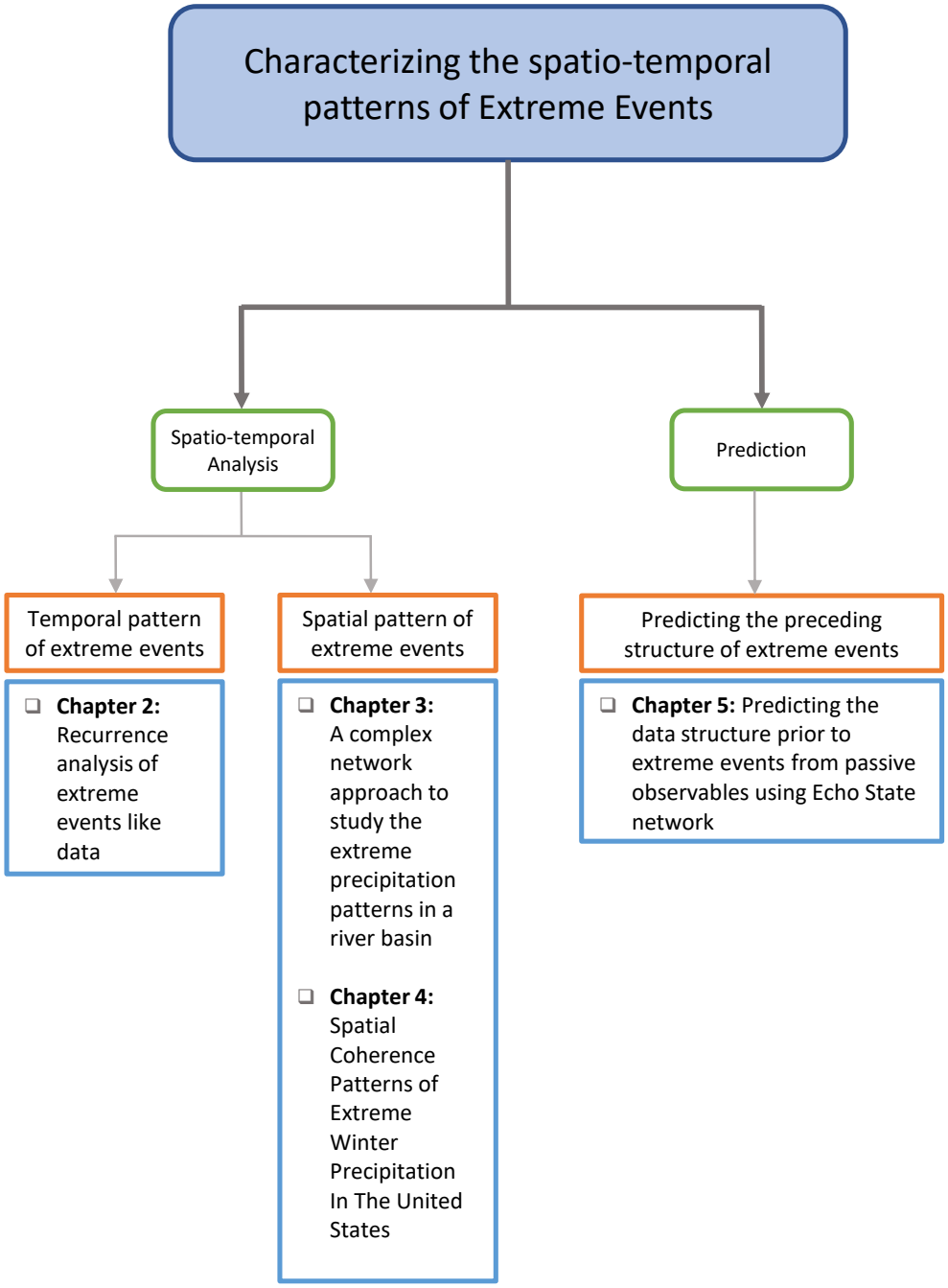
Figure 1.6: Schematic of an Echo State Network. u is the input signal which is fed to the reservoir, W_r, and each node inside the the reservoir consists of a time-dependent state. W_{in} and W_{out} are the input and output weights. The input signal (u) combines with the reservoir state to form the output signal (y).

Recently, machine Learning (ML) based algorithms have gained significant attention in the analysis and forecast of the future states of a complex system (Tang et al., 2020). It has shown considerable potential to analyze and predict EEs in a chaotic system (Doan et al., 2021; Lellep et al., 2020a; Qi and Majda, 2020a). The characteristics of EEs which I learned from my previous investigations, inspired me to delve deeper and develop a predictive framework for EEs using Echo state network, a kind of Recurrent Neural Network based on *reservoir computing*. Reservoir computing (RC) has emerged as a popular machine learning method due its simple architecture, low computational cost, and high predictive skill (Jaeger, 2001; Lukoševičius and Jaeger, 2009; Lukoševičius, 2012). RCs have been successfully used in the prediction of chaotic signals (Jaeger and Haas, 2004a; Pathak et al., 2017, 2018), in speech recognition (Verstraeten et al., 2006), in the prediction of sea surface temperature (Walleshauser and Bollt, 2022), the onset of Indian monsoon (Mitsui and Boers, 2021), and epidemic spreading (Ghosh et al., 2021), etc. *Echo state network* (ESN) is one such approach in RC paradigm which has been a key ingredient in advancing RC and remains a popular choice (Lukoševičius and Jaeger, 2009; Lukoševičius, 2012). Fig.1.6 represents a schematic of ESN. An input signal $\mathbf{u(t)} = (u_1(t), u_2(t), u_3(t), ..., u_k(t))^T$ is fed into the reservoir $W_r^{N\times N}$. Each node has a time

dependent state $\mathbf{r(t)} = r_1(t), r_2(t), r_3(t), ..., r_N(t))^T$ which evolves with time (Lukoševičius, 2012). In ESN, the input ($\mathbf{W}_{in}$), and the reservoir weight ($\mathbf{W}_r$) are chosen randomly. We only tune the output weight ($\mathbf{W}_{out}$).

In Chapter 5, we explain the ESN architecture in detail and device a method based on it to predict preceding structure of EEs. Furthermore, I also found that the efficiency of prediction depends on the intensity or height of the event (Zamora-Munt et al., 2013; Bonatto et al., 2011). This has been observed earlier in case of rogue waves in semiconductor lasers, which are rare ultra-high pulses. Such rogue waves are catastrophic extreme events that occur in many natural systems, such as freak waves in oceans. The dependence of quality of prediction on the magnitude of extreme events has been associated with an increased signal-to noise ratio in case of larger events, leading to better predictability. However, this alone is not sufficient to explain the phenomena under all circumstances. Some previous investigations of EEs in both low-dimensional chaotic systems (Granese et al., 2016) and high-dimensional turbulent systems (Birkholz et al., 2015) found that regular patterns precede the occurrence of an EE because a trajectory has to visit a narrow region in phase space to trigger an EE. To understand the mechanism causing enhanced predictability in a quantitative way in Chapter 5, we propose a metric to measure the phase coherence between the input time signals prior to the occurrence of an EE.

The flowchart below provides a concise outline of my thesis which shows the logical progression of my PhD work:



1.3 Authors contributions

P1. **Banerjee, A.**, Goswami, B., Hirata, Y., Eroglu, D., Merz, B., Kurths, J., and Marwan, N.: Recurrence analysis of extreme event-like data, Nonlin. Processes Geophys., 28, 213-229, https://doi.org/10.5194/npg-28-213-2021, 2021.

AB's contribution: Formulated the study, carried out the analysis and prepared all the figures, and took the lead in writing the first draft of the manuscript. All authors discussed the results, drew conclusions, and provided critical feedback to the manuscript text. All authors were involved in deciding the context and interpretation of the main results.

P2. Agarwal, A., Guntu, R K., **Banerjee, A.**, Gadhawe, M A., and Marwan, N., A complex network approach to study the extreme precipitation patterns in a river basin, Chaos 32, 013113 (2022) https://doi.org/10.1063/5.0072520

AB's contribution: Formulated a part of the study, namely, comparison of edit distance and event synchronization, and conducted the numerical analysis and preparation of the figures related to the same. Additionally involved in writing the original draft of the manuscript and editing and reviewing. RG did the formulation and related numerical analysis of the remaining part; RG was also involved in writing the original draft along with reviewing and editing. All authors discussed the results, drew conclusions, and provided critical feedback to the manuscript text.

P3. **Banerjee, A.**, Kemter, M. , Goswami, B. et al.: Spatial Coherence Patterns of Extreme Winter Precipitation In The United States, Submitted to *Theoretical and Applied Climatology*, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-1243541/v1]

AB's contribution: Formulated the study, collected and prepossessed the datasets carried out the analysis and prepared all the figures, and took the lead in writing the first draft of the manuscript. All authors discussed the results, drew conclusions, and provided critical feedback to the manuscript text.

P4. **Banerjee, A.**, Mishra, A. , Dana, S K. et al.: Prediction of extreme events from passive observables using Echo state network, Submitted to *Frontiers in Applied Mathematics and Statistics*

AB's contribution: Conceived the problem and formulated the study, carried out the numerical analysis and prepared all the figures, and took the lead in writing the first draft of the manuscript. All authors discussed the results, drew conclusions, and provided critical feedback to the manuscript text.

Chapter 2

Recurrence analysis of extreme event-like data

Abstract

The identification of recurrences at various time scales in extreme event-like time series is challenging because of the rare occurrence of events which are separated by large temporal gaps. Most of the existing time series analysis techniques cannot be used to analyze extreme event-like time series in its unaltered form. The study of the system dynamics by reconstruction of the phase space using the standard delay embedding method is not directly applicable to event-like time series as it assumes a Euclidean notion of distance between states in the phase space. The *edit distance* method is a novel approach that uses the point-process nature of events. We propose a modification of edit distance to analyze the dynamics of extreme event-like time series by incorporating a nonlinear function which takes into account the sparse distribution of extreme events and utilizes the physical significance of their temporal pattern. We apply the modified edit distance method to event-like data generated from point process as well as flood event series constructed from discharge data of the Mississippi River in USA, and compute their recurrence plots. From the recurrence analysis, we are able to quantify the deterministic properties of extreme event-like data. We also show that there is a significant serial dependency in the flood time series by using the random shuffle surrogate method.[1]

[1]Originally published as (P1): Banerjee, A., Goswami, B., Hirata, Y., Eroglu, D., Merz, B., Kurths, J., and Marwan, N.: Recurrence analysis of extreme event-like data, Nonlin. Processes Geophys., 28, 213–229. https://doi.org/10.5194/npg-28-213-2021

2.1 Introduction

One of the main challenges of the society is to understand and manage natural disasters, such as earthquakes, tsunamis, and floods, which often lead to big loss of economic assets and even lives. Flooding is an important example with a high societal relevance and affects more people globally than any other natural hazard. Globally, the expected annual costs of floods have been estimated to more than US$ 100 billion (Desai et al., 2015). Furthermore, climate change projections point to an increasing flood risk. Direct flood damages could rise by 160 to 240% and human losses by 70 to 83% in a $1.5°C$ warmer world (Dottori et al., 2018). Climate change has already influenced river flood magnitudes (Blöschl et al., 2019), and has been related to increases in the intensity and frequency of heavy precipitation events aggregating flash flood and river flood risk (Donat et al., 2016; Kemter et al., 2020). Natural climate variability at different time scales may lead to flood-rich and flood-poor periods (Merz et al., 2018). In addition, human interventions in river systems and catchments also heavily influence flood magnitudes and frequencies (Hall et al., 2014). Since floods can massively affect life quality of our societies, it is desirable to understand the underlying dynamics and, thus, put forward precautionary measures to avert potential disasters.

The occurrence of extreme events is not random but rather a manifestation of complex dynamics, and such events tend to have long-term correlation. Comparing data of extreme events with other non-event data using standard methods (linear/nonlinear) is problematic, because the temporal sampling differs largely (time points of events vs. continuous sampling). Furthermore, linear methods such as the Fourier transform (Bloomfield, 2004) and wavelet analysis (Percival and Walden, 2000) are often insufficient to capture the full range of dynamics occurring due to the underlying nonlinearities (Marwan, 2019). Hence, defining a principled nonlinear method is necessary for the analysis of extreme event time series, in particular when the correlation or coupling between several variables is investigated.

Out of the various approaches to study nonlinear dynamical systems (Bradley and Kantz, 2015), the reconstruction of the phase space using delay coordinates (Takens, 1981) is a widely used method that allows us to estimate dynamical invariants by constructing a topologically equivalent dynamical trajectory of the original (often high-dimensional and unknown) dynamics from the measured (scalar) time series. In the delay coordinate approach, the distance between states in the phase space plays a pivotal role in describing the underlying dynamics of a system. After reconstructing the dynamical trajectory, we can extract further information about the dynamics of a system encoded in the evolution of the distances between the trajectories, e.g., through recurrence plots (Marwan et al., 2007), correlation dimension (Grassberger and Procaccia, 1983a), Kolmogorov entropy (Grassberger and Procaccia, 1983b), or Lyapunov exponents (Wolf et al., 1985).

Although there are powerful techniques based on phase space reconstruction of a wide range of nonlinear dynamical processes, they are not directly applicable to event-like time series. Extreme events such as flood, earthquakes, or solar flares are known to have long-term correlations (Jentsch et al., 2006). However, capturing the correlations of extreme events using such methods is difficult as the phase space reconstruction and the Euclidean distance for measuring the distances of states are not suitable for event-like time series because, by definition, extreme events are small in number and are separated by large temporal gaps. In this case, it becomes necessary to define an appropriate distance measure that can help analyze the dynamics of extreme event-like time series. Event-like time series can be analyzed in their unaltered form by considering a time series of discrete events as being generated by a point process. Victor and Purpura (1997) presented a new distance metric to calculate a distance between two spike trains (binary event sequences) as a measure of similarity. Hirata and Aihara (2009) extended this idea called *edit distance* for converting a spike train into time series. The method has also

been adopted to measure a distance between marked point processes to analyze foreign currencies (Suzuki et al., 2010), and irregularly sampled palaeoclimate data (Eroglu et al., 2016). Although the existing definition of *edit distance* is quite suitable to measure similarity between event series, it introduces a bias (discussed in Sec. 2.3), when there are large gaps in the data as in extreme event time series. Additionally, the method depends on multiple parameters and often it is difficult to associate a physical meaning to them. This further complicates the parameterization of the method in case of extreme events.

In this study, we propose a modification of the edit distance metric for analyzing extreme event-like time series. The proposed extension allows to consider the shifting parameter of the edit distance metric in terms of a temporal delay which can be physically interpreted as a tolerance introduced to deal with the quasi-periodic nature of a real world extreme event time series. We demonstrate the efficacy of the proposed modified edit distance measure by employing recurrence plots and their quantification for characterizing the dynamics of flood time series from the Mississippi river in the United States. A flood time series shows a complex time-varying behavior. Moreover, flood generation is often characterized by nonlinear catchment response to precipitation input or antecedent catchment state (Schröter et al., 2015), requiring methods able to deal with nonstationarity and nonlinearity. By using the random shuffle surrogate method, we show that there is a significant serial dependency in the flood events.

2.2 Edit Distance for event-like time series

Distance measurements between two data points play an important role for many time series analysis methods, for example, in recurrence quantification analysis (RQA) (Marwan et al., 2007), estimation of the maximum Lyapunov exponent (Rosenstein et al., 1993), scale-dependent correlations (Rodó and Rodríguez-Arias, 2006), data classification (Sakoe and Chiba, 1978), and correlation dimension estimation (Grassberger, 1983). In case of regularly sampled data, the Euclidean distance is often used. However, for event-like data where big gaps between events are common, this approach is not directly applicable.

Event-like time series can be analyzed in their unaltered form by considering a time series of discrete events as being generated by a point process. Victor and Purpura (1997) presented a specific distance metric to calculate a distance between two spike trains (binary event sequences) as a measure of similarity. Hirata and Aihara (2009) extended this idea for analysing event-like time series and named it *edit-distance*. Ozken et al. (2015) suggested a novel interpolation scheme for irregular time series based on the edit distance, and Ozken et al. (2018) extended this approach to perform recurrence analysis for irregularly sampled data.

In order to apply the edit distance as a distance measure for, e.g., recurrence analysis, the whole time series is divided into small, possibly overlapping, segments (windows), which should contain some data points (Fig. 2.1). The aim is to determine a distance measure between every pair of segments. As a distance, we use the effort of transforming one segment into another using a certain set of operations. For this a combination of three elementary operations is required: (1) delete or (2) insert a data point, and (3) shifting a data point to a different

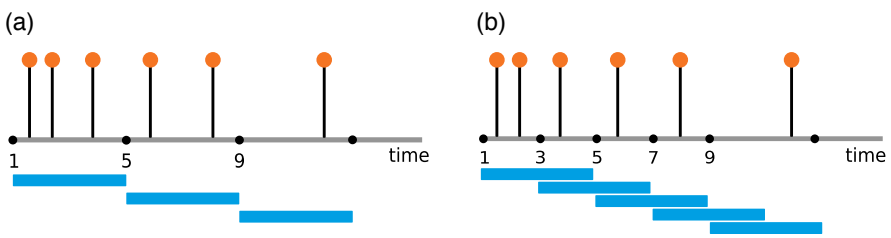


Figure 2.1: (a) Non overlapping windows of window size 4; (b) overlapping windows with 25% sharing.

point in time; each of these operations is assigned a cost. We obtain the distance measure by minimizing the total cost of this transformation. In Fig. 2.2, we show an illustrative example of how the elementary operations transform an arbitrary segment S_a to a segment S_b.

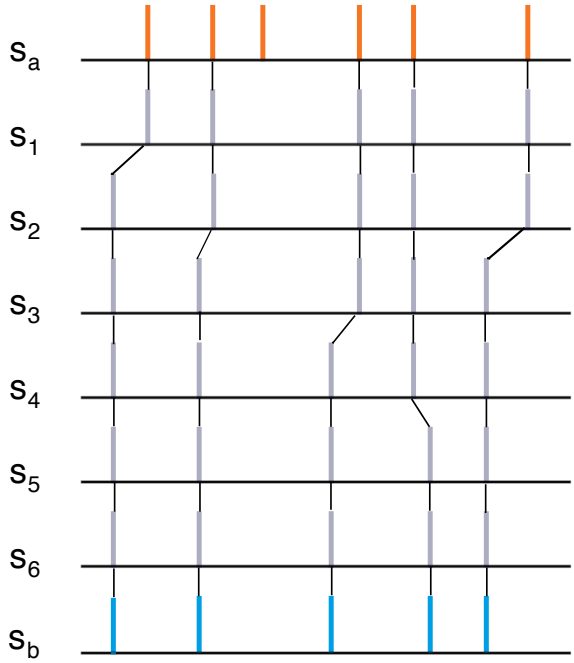


Figure 2.2: A minimum-cost path is shown to transform S_a to S_b.

For the edit distance, we use the sum of the costs of the three operations. The cost function is

$$P(C) = \Lambda_s(\mid I \mid + \mid J \mid -2 \mid C \mid) + \sum_{(\alpha,\beta)\in C} \Lambda_0 \|t_a(\alpha) - t_b(\beta)\|,$$

where α and β are events in segments S_a and S_b occurring at times $t_a(\alpha)$ and $t_b(\beta)$; C is the set containing all such pairs (α, β) from the two segments; I and J are the sets of indices of events in S_a and S_b; $\mid I \mid$, $\mid J \mid$ and $\mid C \mid$ are the cardinalities of I, J and C; Λ_s is the cost of deletion/insertion and Λ_0 the cost assigned for shifting events in time. The first summand in the cost function deals with deletion/insertion and the second summand (the summation) deals with shifting of the pairs $(\alpha, \beta) \in C$.

The distance D is the minimum cost needed to transform the event sequence in S_a to the event sequence in S_b

$$\begin{aligned} D(S_a, S_b) &= \min_C P(C) \\ &= \min_C \left\{ \Lambda_s(\mid I \mid + \mid J \mid -2 \mid C \mid) + \sum_{(\alpha,\beta)\in C} \Lambda_0 \|t_a(\alpha) - t_b(\beta)\| \right\}. \end{aligned} \tag{2.1}$$

The operation costs Λ_s and Λ_0 are defined as

$$\Lambda_s = \text{const.}, \text{ and} \tag{2.2a}$$

$$\Lambda_0 = \frac{M}{\text{total time}}, \tag{2.2b}$$

where M is the number of all events in the full time series. Originally, Λ_s was chosen as a constant value 1 (Victor and Purpura, 1997). Later, Λ_s was optimized in a range $[0, 4]$ (Ozken et al., 2018). The units of the parameters are $(\text{time})^{-1}$ for Λ_0 and unit-less for Λ_s.

The treatment of an event series as point process makes the edit distance measure a good starting point for defining a distance between segments of an extreme event time series. However, the existing form of the edit distance has a linear dependency on the difference between the occurrences of events which is inappropriate for an extreme event time series, as the rare occurrence leads to large gaps between events. Also, as already mentioned, the existing method depends on a number of parameters. Therefore, we suggest a modification of the cost function to address these two concerns.

2.3 Modified Edit Distance

The cost of transformation between related events (e.g. events belonging to the same climatological phenomenon in a climate event series) should be lower than that between independent, unrelated events. Hence, the shifting operation should be a more likely a choice for comparison between segments if the events in each segment are related or belong to the same phenomenon. Consequently, deletion/insertion as a choice of operation for transformation tends to be associated with unrelated events.

Now, we consider two event sequences S_a and S_b, where each of them has only one event at times t_a and t_b, respectively (Fig. 2.3). We want to transform segment S_a to S_b by using either deletion/insertion or shifting. The path for shifting the solitary spike has the cost $\Lambda_0 \Delta t$ where $\Delta t = \mid t_a - t_b \mid$, i.e., it grows linearly with the distance between the two events. On the other hand, the cost to delete the event at t_a from S_a and insert it at time t_b in S_a for it to resemble S_b has cost 2, as the single operation cost for deletion and insertion is each 1. So the shifting operation will take place only as long as the time difference between both events is smaller then $\Delta t < \frac{2}{\Lambda_0}$.

The above condition has two limiting cases which need to be considered for an unbiased choice between shifting and deletion/insertion – when the cost of shifting is too low or when it is too high. The former case arises for an extreme event time series where the number of extreme events $M \ll$ total length of the time series, i.e., $\Lambda_0 \to 0$, Eq. (2.2b). This assigns a low cost for the transformation due to a biased choice of shifting over deletion/insertion for largely separated events which may be independent.

The other limiting case, when the cost of shifting per unit time, Λ_0, is moderately high, leads to a biased preference for the deletion/insertion operation over shifting depending on how high Λ_0 is, because the cost of shifting increases linearly with the distance between two events according to the definition in Eq. (2.1). In this case, related events separated by a relatively small gap may be considered as independent as the shifting cost may exceed the deletion/insertion cost because of higher Λ_0. The following example of a real extreme event series illustrates how this biased preference can lead to erroneous results. In regions with

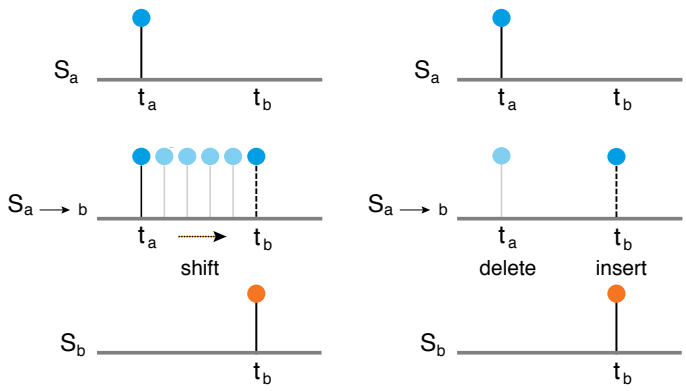


Figure 2.3: Figures illustrating the operation of shifting (left) and deletion/insertion (right).

seasonal precipitation regimes, high precipitation events are more similar to those of the same season of another year compared to other seasons. However, the exact time of occurrence of the extreme precipitation events varies from year to year, i.e., even if the events in each segment are related, there might be a certain time delay between the events (Fig. 2.4). Now, a linear dependency of the shifting cost on the difference between times of events, Eq. (2.1), does not allow us to consider small delays between potentially similar events. In addition, a not so high value of Λ_0 could increase the cost of shifting for small delays higher than the cost of deletion/insertion, implying a lower similarity between the segments.

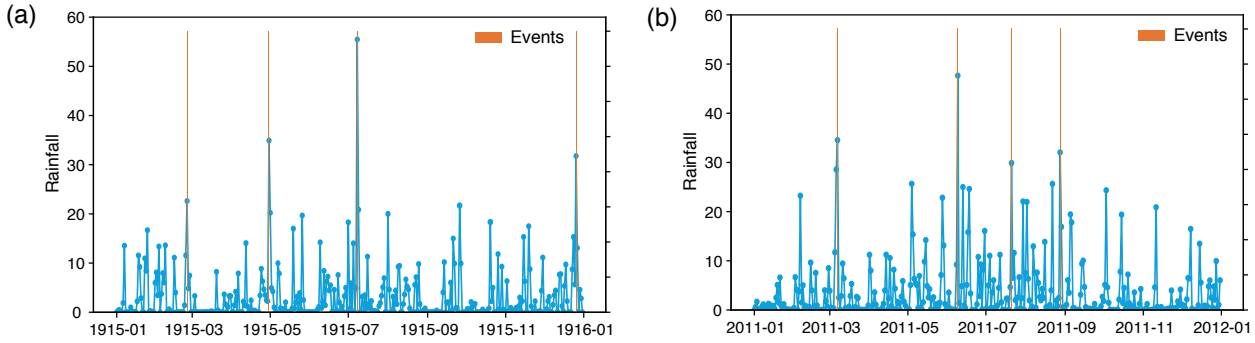


Figure 2.4: Daily precipitation time series of the Aroostook River catchment near Masardis, Maine in the United States for the years (a) 1915 and (b) 2011; red bars indicate extreme rainfall events (4 events per year), usually occurring in spring and summer, whereas in autumn extreme rainfall events usually do not occur.

As the maximum cost for shifting is limited by the cost for deletion/insertion, we need to lower the cost of shifting to get a higher similarity between the segments for the above case. This modification is done by considering the temporal distance Δt at which the maximum cost of shifting occurs as a delay between the two events.

Selecting a certain temporal delay is relevant for climate time series, such as precipitation, or hydrological variables, like discharge, which are of an event-like nature, where the synchronization between extreme events (Malik et al., 2010, 2012; Boers et al., 2013, 2014d; Ozturk et al., 2018) at different geographical locations and recurrence of extreme events for the same location are of interest.

The above discussion illustrates the importance of an appropriate choice of the cost of shifting per unit time, Λ_0, based on properties of the event series, such as time series length, total number of events, and event rate. Thus it is more reasonable to optimize Λ_0 as opposed to Λ_s. Victor and Purpura (1997) suggested generalizations of the cost assigned to finite translations, such as modifying the simple $\Lambda_0 \Delta t$ to more general functions, satisfying the triangle inequality.

In view of comparing two events under a predefined delay, we suggest to replace the linear cost function for shifting, Eq. (2.1), by a nonlinear cost function which allows a temporal tolerance and also ensures a smooth change from the cost function for shifting to the cost for deletion/insertion (Fig. 2.5). We propose to use the Logistic function (Cramer, 2002) as the cost for shifting:

$$f(t) = \frac{\Lambda_s}{1 + e^{-k(t-\tau)}}, \tag{2.3}$$

where

- τ is the chosen time interval between events marking the transition of the function from exponential growth (for closely-spaced events) to bounded exponential growth (for events separated by large time gaps) (Fig. 2.6a),
- Λ_s is the maximum cost, and

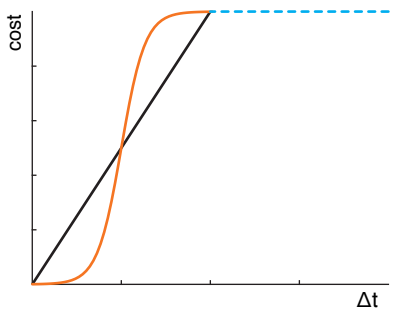


Figure 2.5: Linear cost function for shifting (black) and alternative, nonlinear cost function (orange).

- k is a parameter that affects the rate of the exponential growth, in this study we choose $k = 1$. For $k = \infty$, the Logistic function becomes a step function.

The Logistic function (Eq.(2.3)) is used for modeling the population growth in an area with limited carrying capacity.

The sigmoid nature of the Logistic function is apt for representing the cost of shifting with low values for closely-spaced events and which tends to saturate for events separated by a time delay greater than a certain τ.

A step function (e.g., Heaviside function) would be another choice for the cost function according to whether shifting is chosen below a certain delay and deletion/insertion is chosen above it (Fig. 2.6b):

$$f(t) = \begin{cases} \Lambda_0 & \Delta t \leq \tau \\ \Lambda_s & \Delta t > \tau \end{cases} \tag{2.4}$$

where Λ_0 is the cost for shifting, according to Eq. (2.2b), and τ is the given delay choice which decides between shifting and deletion/insertion. However, such a cost function maintains a constant value irrespective of the event-spacing and switches to a different value only at the delay threshold. Therefore, the logistic function, as a smooth approximation of the step function, is a preferable choice.

The parameter optimization problem of the modified edit distance equation now becomes a problem of solving a single linear equation with two unknown variables – the coefficient related to the cost of shifting events in time and Λ_s. Since, in our method, we choose to optimize the former by using the Logistic function, we keep Λ_s constant.

This constant can be absorbed in the optimization function of the first term, and therefore, the coefficient of the second term related to the cost of addition/ deletion.

For a particular pair of events (α, β), we can arbitrarily set the maximum cost Λ_s for shifting or deletion/insertion to be 1, because it is the only free cost parameter (we have no Λ_k because of neglecting the amplitudes). Thus, the cost for transforming one segment to another using the *modified edit distance* (mED) is defined as

$$P(C) = (\mid I \mid + \mid J \mid -2 \mid C \mid) + \sum_{(\alpha,\beta)\in C} \left\{ \frac{1}{1 + \mathrm{e}^{-k(\|t_a(\alpha) - t_b(\beta)\| - \tau)}} \right\}$$

and the corresponding distance function to calculate the distance between two segments S_a and S_b is

$$D(S_a, S_b) = \min_C \left\{ (\mid I \mid + \mid J \mid -2 \mid C \mid) + \sum_{(\alpha,\beta)\in C} \left\{ \frac{1}{1 + \mathrm{e}^{-k(\|t_a(\alpha) - t_b(\beta)\| - \tau)}} \right\} \right\}. \tag{2.5}$$

This new definition of cost depends only on the parameter τ, which can be interpreted in the sense of a delay between events. Moreover, this definition holds the triangular inequality when

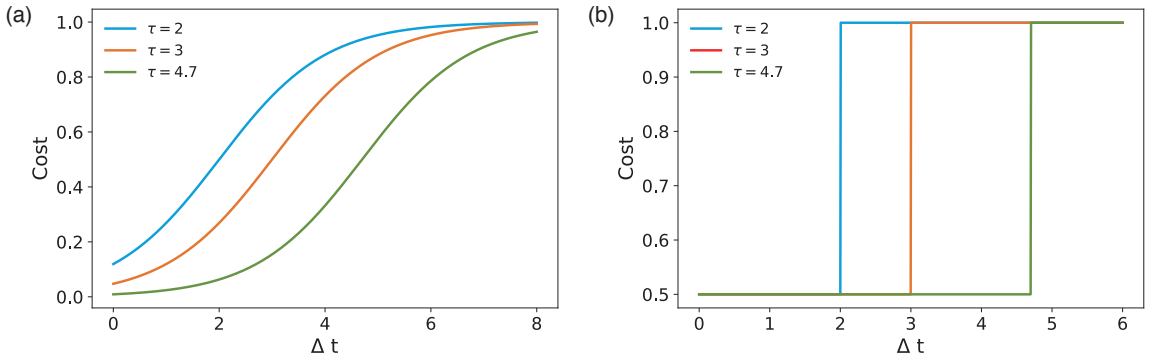


Figure 2.6: Variation of the cost contributed by shifting an event by Δt for different parameter values $\tau = 2, 3,$ and 4.7 using (a) the logistic function with $k = 1$ and (b) the Heaviside function.

τ satisfies a certain condition (see the Appendix for proof).

In the next section we use the modified version of *edit distance* to perform recurrence analysis of extreme event-like data.

2.4 Recurrence plots

A recurrence plot (RP) is a visualization of the recurrences of states of a dynamical system, capturing the essential features of the underlying dynamics into a single image. The quantification of the patterns in a RP by various measures of complexity provides further (quantitative) insights into the system's dynamics (recurrence quantification analysis, RQA) (Marwan et al., 2007). RQA was designed to complement the nonlinearity measures such as the Lyapunov exponent, Kolmogorov entropy, information dimension, and correlation dimension (Kantz, 1994). RP-based techniques have been used in many real world problems from various disciplines. In the field of finance, Strozzi et al. (2002) studied RQA measures for high frequency currency exchange data. In astrophysics, Stangalini et al. (2017) applied RQA measures to detect dynamical transitions in solar activity in the last 150 years. It has also been used to classify dynamical systems (Corso et al., 2018), or to detect regime changes (Marwan et al., 2009; Eroglu et al., 2016; Trauth et al., 2019) and has many applications in Earth science (Marwan et al., 2003; Chelidze and Matcharashvili, 2007), econophysics (Kyrtsou and Vorlow, 2005; Crowley and Schultz, 2010), physiology (Webber, Jr. and Zbilut, 1994; Zbilut et al., 2002; Marwan et al., 2002; Schinkel et al., 2009), and engineering (Gao et al., 2013; Oberst and Tuttle, 2018).

Consider a time series that encodes N measured states of a dynamical system $\{\vec{x}_i \in M\}_{i=1}^{N}$ on space M. A state of this system is said to be recurrent if it falls into a certain ε-neighborhood of another state. Given a distance function $D : M \times M \to \{0\} \cup \mathbf{R}^+$, for a given trajectory $\vec{x}_i$, the *recurrence matrix* of the system is defined as:

$$R_{i,j}(\varepsilon) = \begin{cases} 1, & \text{if } D(\vec{x}_i, \vec{x}_j) < \varepsilon, \\ 0, & \text{otherwise.} \end{cases} \tag{2.6}$$

In a RP, when $R_{i,j}(\varepsilon) = 1$ a point is plotted at (i, j), otherwise nothing is plotted. Different classes of dynamics result in different patterns in their respective RPs (Marwan et al., 2007).

The RP contains a main diagonal line, called the line of identity (LOI), corresponding to the recurrence of a state with itself. The RP is symmetrical about the LOI when D is a metric (e.g., a symmetrical norm). We are interested in the line structures in RP as they capture several aspects of the underlying dynamical behaviour of a system. For instance, long, continuous lines parallel to the LOI denote (pseudo-)periodic behavior, whereas short, discontinuous diagonal lines are indicative of a chaotic system.

In order to incorporate mED in RP, we divide the time series into small segments and compute the distance between these segments; the time indices are the centre points of these segments. We can now define the RP in terms of the distance between the segments calculated by mED:

$$R_{i,j}(\varepsilon) = \Theta(\varepsilon - D(S_i, S_j)) \quad i, j = 1, 2, ..., \omega, \tag{2.7}$$

where ω is the number of segments and the size of the RP is $\omega \times \omega$.

One of the most important measures of RQA is the determinism (DET), based on the diagonal line structures in the RP. The diagonal lines indicate those time periods where two branches of the phase space trajectory evolve parallel to each other in the phase space. The frequency distribution $P(l)$ of the lengths of the diagonal lines is directly connected to the dynamics of the system (Marwan et al., 2007).

$$DET = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{l=1}^{N} lP(l)}. \tag{2.8}$$

In our study we choose $l_{min} = 2$. RPs of stochastic processes mainly contain single points, resulting in low DET values, where RPs of deterministic processes contain long diagonal lines, resulting in high DET values.

In this work, we are interested in the deterministic nature of flood event time series. For this purpose, we focus on the DET measure.

2.5 Choice of window size

We divide the time series into segments/windows. If adjacent windows overlap, we call it overlapping sliding window, else, non-overlapping window (Fig. 2.1). Sliding window techniques are widely used for signal processing (Bastiaans, 1985), activity recognition processes (Dehghani et al., 2019). The window size should be selected properly, i.e. each window should contain enough data points to be differentiable from similar movements. Consider a time series of data $x_i \in \mathbb{R}$ at times $t_i \in \mathbb{N}$. For generalization we consider a constant sampling rate, i.e.,

$$\Delta T = t_{i+1} - t_i = \text{const.} \ \forall i \in \mathbb{N}.$$

Each window consists of n ($n \in \mathbb{N}, n > 1$) data points, so the window size is $w = n\Delta T$. For overlapping windows, a fraction of the data is shared between consecutive windows, denoted by $L \in \{1, 2, 3, .., n-1\}$ as the number of data points within the overlapping range, where $L = \varnothing$ signifies the non-overlapping case. The overlapping range is $O_L = L\Delta T$ and in terms of percentage $O_L(\%) = \frac{L}{n}100\%$.

Because mED focuses on event-like data, finding an optimum window size is important. An inappropriate window size may lead to missing events in windows ("empty windows") due to the sparse occurrence of events. Here we propose some criteria for choosing the optimum window size:

- To avoid empty windows we try to fix the number of events (n) in each window. Here we choose the window size as 1 year, as most climate phenomena such as flood events exhibit annual periodic behaviour and therefore we get similar number of events each year.
- In case we need to choose a larger window size to have enough data points in each window, the number of segments decreases if the windows do not overlap which in turn reduces the dimension of the recurrence matrix. The underlying dynamical behaviour of the system will not be completely captured by the resulting coarse-grained recurrence plot.

Overlapping windows alleviate this problem by increasing the number of windows.

In our work, we use a certain percentage of overlapping.

- An inappropriate window size can lead to the aliasing effect (Fig. 2.11). As a result different signals would be indistinguishable and we might loose important transitions.

2.6 Applications of modified edit distance

We apply mED first to generated data and then to real-world flood observations to understand how well our method can identify recurrences in extreme event-like time series.

2.6.1 Numerically generated event-series using Poisson process

The Poisson process is used as a natural model in numerous disciplines such as astronomy (Babu and Feigelson, 1996), biology (Othmer et al., 1988), ecology (Thompson, 1955), geology (Connor and Hill, 1995), or trends in flood occurrence (Swierczynski et al., 2013). It is not only used to model many real-world phenomena but also allows for a tractable mathematical analysis.

Consider $N(t)$ to be a stochastic counting process which represents the number of events above some specified base level in the time period $(0, t)$. Suppose the events occur above the base level at a constant rate $\lambda > 0$ (units of 1/time). So, the probability that n events occur in the time between t and $t + s$ is given by (Ross, 1997; Loaiciga and Mariño, 1991)

$$P\{N(t+s) - N(s) = n\} = e^{-\lambda t}\frac{(\lambda t)^n}{n!} \quad (n = 0, 1, 2, \ldots). \tag{2.9}$$

A Poisson process is a set of random events whose stochastic properties do not change with time (stationary) and every event is independent of other events, i.e., the waiting time between events is memory-less (Kampen, 2007).

Here, we study three cases of numerically generated event series which are motivated by the occurrence of natural events. First, we test mED for a simple Poisson process. Each subsequent case tries to capture features of these real world phenomena by adding an element of and memory to the simple Poisson process. We compare the RPs and the structures in the RPs (by considering the RQA measure DET) derived from the standard edit distance and from the modified edit distance.

For ED (Eq. 2.1), the cost parameter for shifting is calculated according to Eq. (2.2a) and $\Lambda_s = 1$, whereas Eq. (2.5) is used for. We choose the upper bound for the range of τ to be less than the mean inter-event time gap for the complete time series. The physical interpretation of this choice is that the temporal tolerance or delay in the arrival of events in a particular season (event cluster) should not only be less than the time period of the seasonal cycle but also relatively less than the length of the season.

The selection of the threshold based on a certain percentile of the distance distribution makes the recurrence quantification more stable (Kraemer et al., 2018). Keeping in mind that this threshold should neither be too high nor too low, we chose it as 8 to 10 percentile of the distance distribution based on a number of trials.

It is expected that the quasi-periodic behaviour of the event series will lead to high DET values. In case of ED, the DET should be constantly high at all τ, as ED does not include the concept of time delay. On the other hand, DET computed using mED should first increase with τ and then slowly decrease.

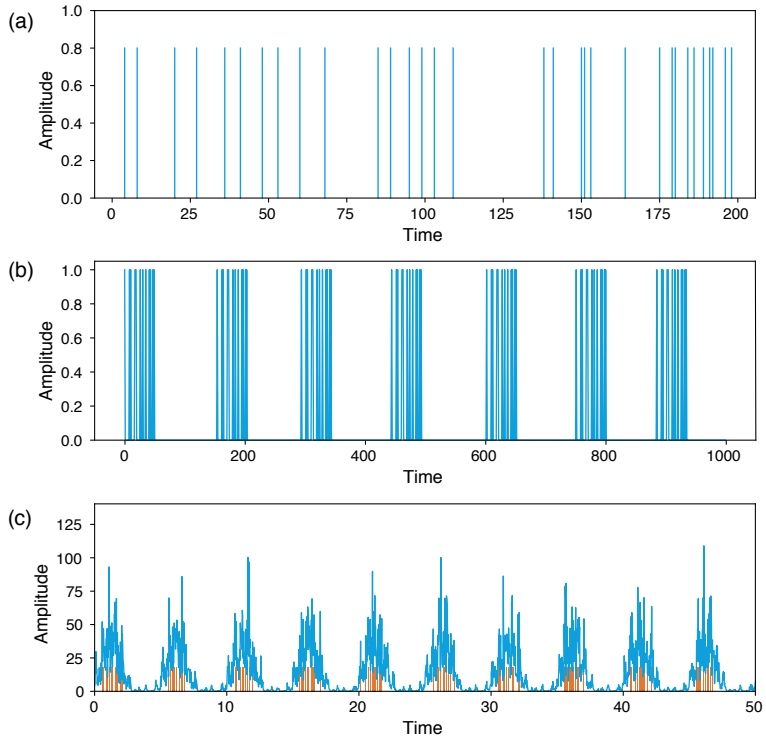


Figure 2.7: Generated time series from a (a) homogeneous Poisson process (total length =10,000), (b) repeating Poisson process (total length=22,442), and (c) Poisson process with periodical forcing to mimic discharge time series (total length 0–500 with 10,000 equally spaced points), small red bars indicate the events.

2.6.1.1 Homogeneous Poisson process

The homogeneous Poisson process models the occurrence of random spaced, i.e. stochastic, events in time, where the average time between the events is known. It is used to model shot noise, radioactive decay, arrival of customers at a store, earthquakes, etc. Here we construct an event series (Fig. 2.7a) from Eq. (2.9) with $\lambda = \frac{1}{5}$. The RP of a realisation of such a homogeneous Poisson process is shown in Fig. 2.8a,b. In a stochastic process, a recurrence of a randomly selected state occurs by chance, resulting in randomly distributed points in the RP. Accordingly, DET has low values (Fig. 2.8c).

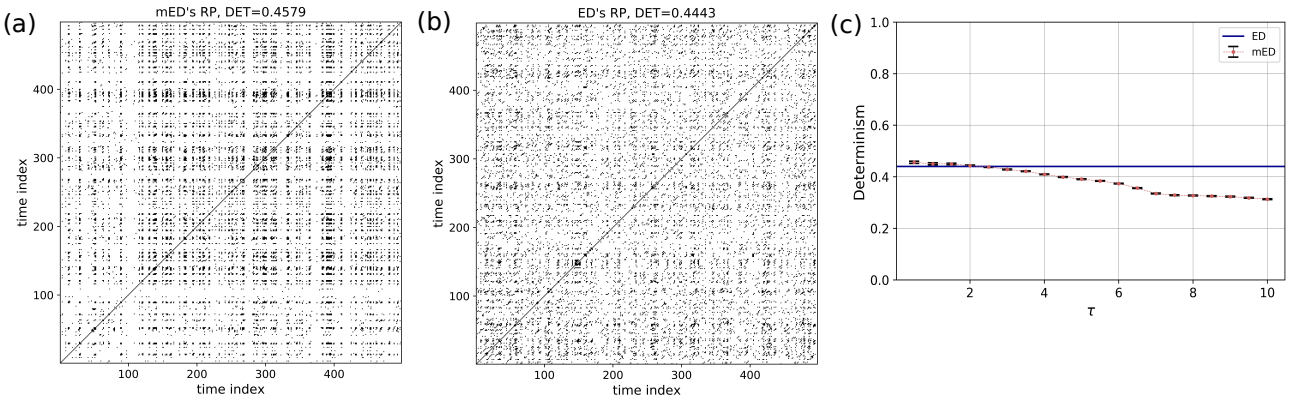


Figure 2.8: RP of a homogeneous Poisson process (Fig. 2.7a) using (a) mED, (b) ED; (c) comparison of DET for 300 realisations using ED (blue horizontal line) and using mED for varying τ in range 0 to 10.

2.6.1.2 Repeating Poisson process

We generate a small segment of a simple Poisson process and repeat the same segment after certain time gaps (Fig. 2.7b). Here, the Poisson process is generated for a segment of length 50 with mean rate of events $\lambda = \frac{1}{5}$. The gap between the event segments is chosen randomly in the range 105 to 115. The RP of this event series is shown in Fig. 2.9. Identical sequences of events result in longer diagonal lines in the RP (with lengths in the order of the event sequences), but the varying time gaps between the event chunks make the diagonal lines discontinuous implying a quasi-periodic behaviour. Quasi-periodicity is often a characteristic of an extreme event time series such as flood events. RP using ED Fig.2.9b contains more short and discontinuous diagonal lines corresponds to less DET value. In Fig.2.9c, we find a certain range where the DET value calculated using mED is higher than ED.

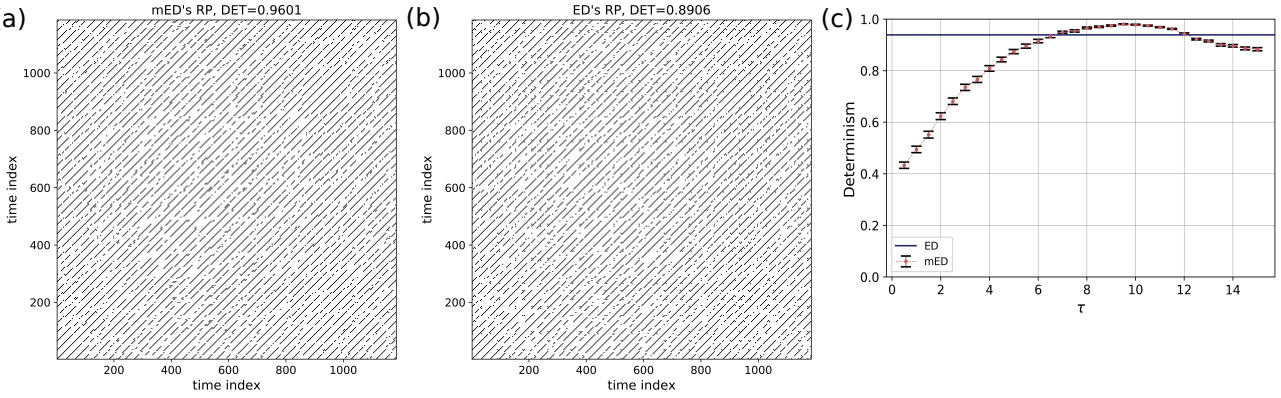


Figure 2.9: RP of a repeating Poisson process (Fig.2.7b) using (a) mED, (b) ED; (c) comparison of DET for 300 realisations using ED (blue horizontal line) and using mED for varying τ in range 0 to 15.

2.6.1.3 Poisson process with periodical forcing

Quasi-periodicity in a time series can be expressed as a sum of harmonics with linearly independent periods. Here we construct a time series with superposition of a slow moving signal $f(T_1) = \sin(2\pi t/T_1)$ and a fast moving signal $f(T_2) = \sin(2\pi t/T_2)$, whereas the events are drawn by a Poisson process (Fig. 2.7c). We select the extreme events by using a certain threshold and take the peak over the threshold (POT). Their occurrences and positions can be seen as the outcome of a point process (Coles, 2001). Such a time series can be used to mimic a streamflow time series which may exhibit periodic behaviour on several scales (annual, seasonal, decadal, etc.). Here the window size is equal to the time period of the slow moving signal. The RP is shown in Fig. 2.10. The periodic occurrences of the event chunks is visible in the RP by the line-wise accumulation of recurrence points with a constant vertical distance (which corresponds to the period). The stochastic nature of the "local" event pattern is visible by the short diagonal lines of varying lengths. As mentioned earlier, when using an improper window size we can lose recurrence points and a smaller number of diagonal lines occurs due to the aliasing effect (Fig. 2.11).

For a certain range of τ, DET for mED (Fig.2.8, Fig.2.9, and Fig.2.10) to be higher than the DET for ED, thus capturing the underlying periodic behaviour better.

2.6.2 Recurrence analysis of flood events

As a case study, we apply our method to the Mississippi river which has a rich history of flood events. Here we are motivated to study the recurring nature of flood events using recurrence

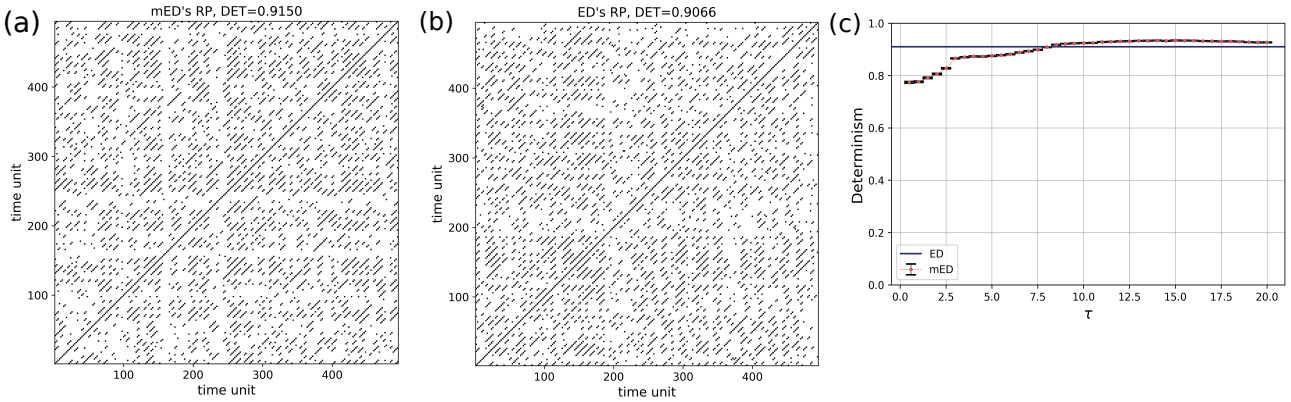


Figure 2.10: RP of Poisson process with periodical forcing (Fig.2.7c) using (a) mED, (b) ED; (c) comparison of DET for 300 realisations using ED (blue horizontal line) and using mED for varying τ in range 0 to 20.

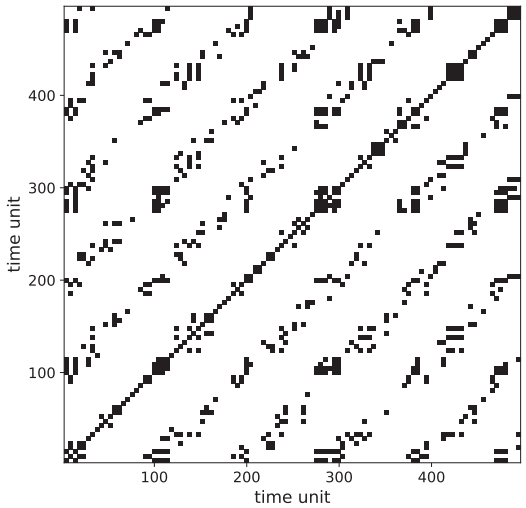


Figure 2.11: RP of Poisson process with periodical forcing (Fig. 2.7c) using mED showing aliasing effect due to improper window size (note that this effect also occurs for the standard ED approach).

plots. Recurrence plot analysis also helps to quantify the serial dependency of flood time series. (Wendi et al., 2019) used RQA to study the similarity of flood events.

We use the mean daily discharge data of the Mississippi River from the Clinton, Iowa station in the USA for the period 1874–2018. The data is obtained from the Global Runoff Data Centre (www.bafg.de/GRDC).

In order to find the events from the time series we follow the procedure below:

1. First, we select events above an arbitrary threshold, say 99^{th} percentile value of the daily discharge time series for a particular year, which gives about 3 to 4 events per year.

2. Next, if several successive days fall above the threshold forming a cluster of events, we pick only the day which has the maximum discharge value and remove the remaining events of the same cluster.

3. Then we lower the threshold by 0.1 percentile (the threshold is lowered from 99^{th} to 90^{th}) and repeat the same procedure as above until we get the desired number of events.

We apply mED for finding recurrence patterns for flood events. To this end, we choose our window size equal to the annual cycle (1 year) with 6 months and 9 months of overlapping and with the cost function of delay, $\tau = 15$, 30, 45 and days. The recurrence threshold $\varepsilon = 8$ percentile of the distance matrix. The recurrence plot of flood events is shown in Fig. 2.12 and 2.13 . We also compute the recurrence plot using ED and calculate DET Fig. 2.14.

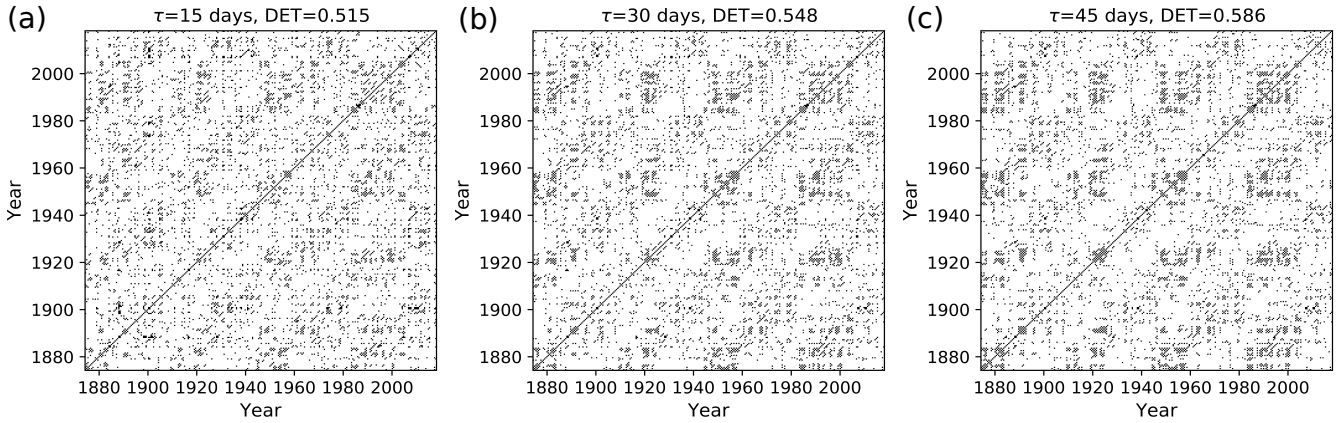


Figure 2.12: Recurrence plot of flood events using mED, window size = 1 year with 6 months overlap for delay τ = (a) 15 days, (b) 30 days, and (c) 45 days.

The black points in the RP (Fig. 2.12, 2.13) denote the segments which have higher similarity and the white points imply less similarity in the occurrence of flood events. The RP of Mississippi floods shown in Fig. 2.13c might look similar to the RP in Fig. 2.8 for events generated by Poisson process. However, the zoomed in image of the RP of Mississippi floods, Fig 2.13d, is very similar to the RP in Fig. 2.10a for events generated by Poisson process with periodical forcing. Diagonal structures are seen in the flood RP (Fig. 2.12, 2.13) denoting an inherent serial dependency in the data as indicated by the DET value. Serial dependence is a property by virtue of which the future depends on the past. When diagonal lines are likely to appear in a RP, the current neighbors tend to be neighbors in the near future, and thus, the corresponding system is said to have serial dependence.

In order to compare the *edit distance* (ED) with the *modified edit distance* (mED) in the study of flood events, we compute the RP and *determinism* DET using both methods. For ED, there is no scope to implement a delay, we get the DET value for a constant temporal gap. In case of mED, because of the predefined delay τ we can study the behaviour at different time scales. In order to do so, we vary the parameter τ in the range from 1 to 60 days. For mED the DET values is slightly lower than ED up to $\tau = 30$ days (Fig. 2.15). But after that, in the range of 30 to 60 days of τ, DET becomes higher for mED. So, for this particular time series we get more deterministic behaviour for delay in range 30 to 60 days.

To determine the statistical significance of the RP analysis, we develop the following statistical test. We use random shuffle surrogates (Scheinkman and LeBaron, 1989) statistical testing. Suzuki et al. (2010) used this method for finding serial dependency on foreign exchange tick data by quantifying the diagonal lines in the RP (DET measure).

We first set a null hypothesis, then we generate a set of random surrogates that preserve the null hypothesis property. After that, we compare the test statistics of the original data with the surrogates data. We can reject the null hypothesis if the test statistics obtained from the surrogate data and the original data is out of the specified range, else we cannot reject the null hypothesis.

The null hypothesis is that there is no serial dependency in the data. Thus, we expect that after each random shuffling the information will be preserved. We create each random surrogate by randomly shuffling rows and columns of a recurrence plot simultaneously. We can reject the null hypothesis if the test statistics of the original data K_0 (DET value) is out of the specified range from the distribution of those surrogate datasets K_S (Theiler et al., 1992). The algorithm works as follows:

1. Calculate the distance matrix and compute the RP for a certain window size.

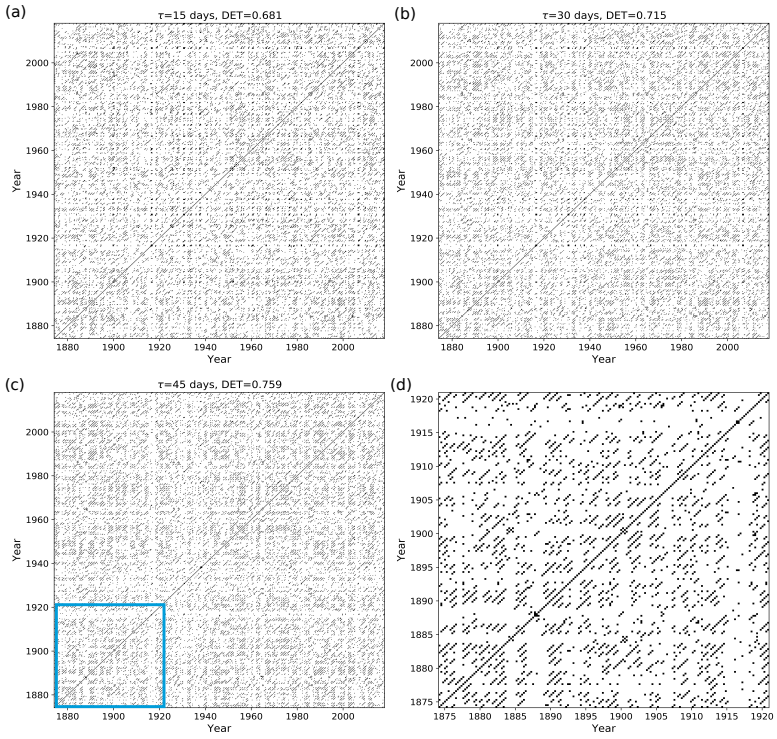


Figure 2.13: Recurrence plot of flood events using mED, window size = 1 year with 9 months overlap for delay τ = (a) 15 days, (b) 30 days, and (c) 45 days; (d) zoom in image of the blue area of (c).

2. Reorder the row and columns randomly to create surrogate recurrence plots.

3. Calculate the DET of the surrogate recurrence plots.

4. Repeat the steps 1–3 and get DET for 50 different surrogate RPs.

We assume the statistics obtained from the surrogates K_s is normally distributed and we consider only one experimental data set. The measure of "significance" as defined by Theiler et al. (1992):

$$\rho = \frac{K_0 - \mu_{K_s}}{\sigma_{K_s}} \tag{2.10}$$

follows a t-distribution, the number of surrogates are $n = 50$ with $(n-1)$ degrees of freedom. μ_{K_s} and σ_{K_s} are the mean and standard deviation of the statistics of surrogate data. The value of ρ has to be compared to the value of the t-distribution that corresponds to the 99^{th} percentile and the mentioned degrees of freedom, i.e., $t_{0.01/2}(49) = 2.68$. If ρ exceeds this value, the null-hypothesis has to be rejected.

We measure ρ for the data generated from a homogeneous Poisson process, Eq. (2.9), and also for the flood event data. First, for the Poisson process the value of $\rho = 1.91, < 2.68$, so we cannot reject the null hypothesis, confirming the missing serial dependency in the homogeneous Poisson process.

For the flood event data, the distribution of the test statistics (DET) is far away from the value of the original data (Fig. 2.16). The ρ value for the flood event RP with 6 months overlapping and $\tau = 15$ days is 23.80 which is much larger than 2.68. Hence, we can reject the null hypothesis at the significance level of $\alpha_0 = 0.01$, getting strong indications that there is serial dependency at high significance level in the occurrence of flood events.

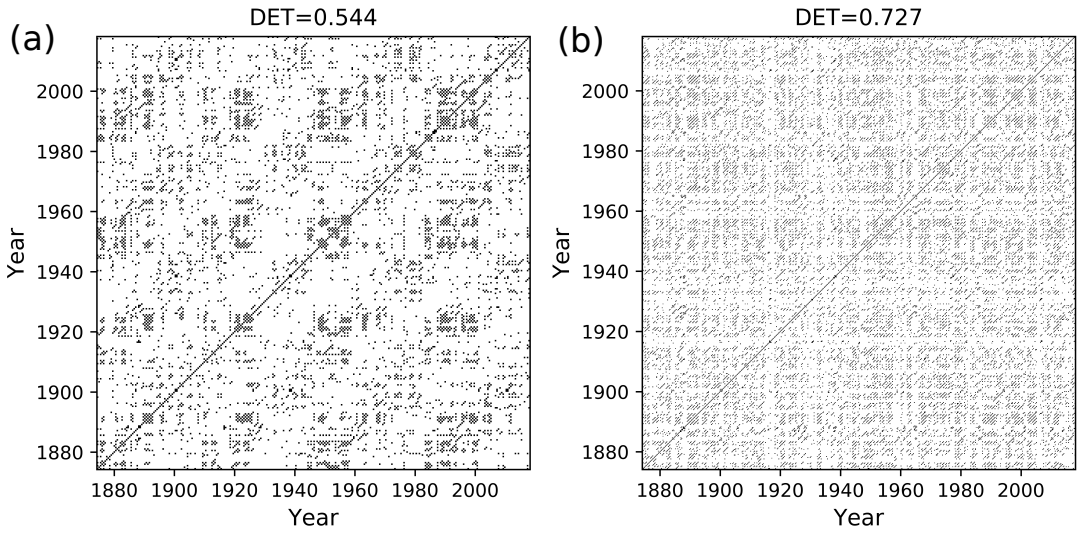


Figure 2.14: Recurrence plot of flood events using ED, window size = 1 year, with (a) 6 months overlapping and (b) 9 months overlapping.

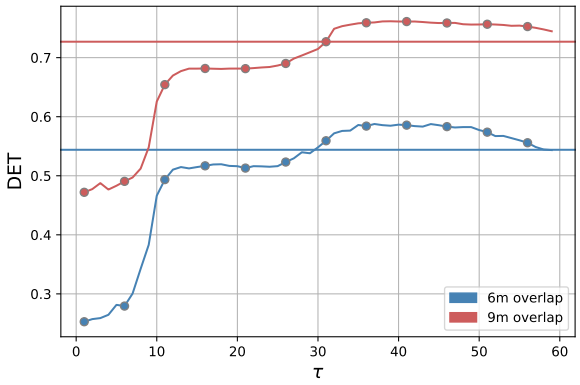


Figure 2.15: Comparison of DET for flood events using ED (horizontal line) and using mED (curved line) for 6 months overlapping (in blue color) and for 9 months overlapping (in red color)

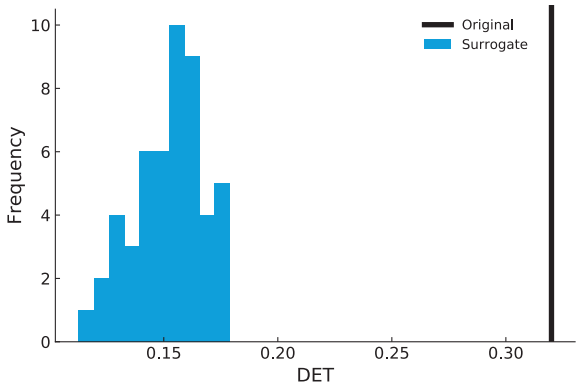


Figure 2.16: Distribution of DET value of surrogate data (blue) and the original DET value (black).

2.7 Conclusion

In this paper, we propose a distance measure for recurrence based analysis of extreme event time series. The proposed measure is based on a modification of the *edit distance* measure proposed by Victor and Purpura (1997). We include the concept of time delay to incorporate the slight variation in the occurrence of recurring events of a real world time series due to changes in seasonal patterns by replacing the linear dependency of the cost of shifting events by a nonlinear dependency using the logistic function. This is substantial improvement over the previous definition of the edit distance as used for the TACTS algorithm proposed by Ozken et al. (2015) as the optimization of Λ_0 is based on the temporal delay between events, which is of physical relevance to the study of extreme events and can be chosen according to the phenomenon being studied. The modified edit distance also reduces the number of independent parameters. We tested mED on prototypical event series generated by a point process (Poisson process). We found that for the quasi-periodic repeating Poisson process event series, the determinism of RP computed using mED varies with temporal delay and is higher than the measured DET value of RP computed using ED for a certain range of delay. We applied the method to study recurrences in the flood events of the Mississippi river. Our analysis revealed deterministic patterns in the occurrence of flood events from RP. Finally, using the random shuffle surrogate method we have shown that the data of the occurrence of flood events has a statistically significant serial dependency.

In this work, we have only considered binary extreme event time series, and ignored the amplitude of events. Next, the mED measure should be further modified by including the cost due to difference in amplitude for applications in real world time series where patterns of intensity of extremes might be of interest. Moreover, an elaborate method to optimize for a proper window size may be devised to capture recurrences of events at different time scales.

Appendix

For deletions and insertions, the new definition inherits the triangle inequality for the original definition of edit distance by Victor and Purpura (1997). Thus, we can show that the new definition of the edit distance preserves the triangle inequality as a total if a shift preserves the triangle inequality. For this sake, considering the following situation is sufficient: Suppose that there are three simple point process each of which has an event; the first point process has an event at $u (u > 0)$; the second point process has an event at $u + s (s > 0)$; the third point process has an event at $u + s + t (t > 0)$. Therefore, the first and the second point processes have an inter-event interval of s; the second and the third point processes have an inter-event interval of t; and the first and the third point processes have an inter-event interval of $s + t$. We assume that $\tau \leq \frac{s+t}{2}$. Then, the distance between the first and the second point processes is

$$D(S_1, S_2) = \frac{1}{1 + e^{-s+\tau}}.$$

Likewise, the distance between the second and the third point processes is

$$D(S_2, S_3) = \frac{1}{1 + e^{-t+\tau}}.$$

The distance between the first and the third point processes is

$$D(S_1, S_3) = \frac{1}{1 + e^{-s-t+\tau}}.$$

Then the following chain of inequalities holds:

$$
\begin{aligned}
&- D(S_1, S_3) + D(S_1, S_2) + D(S_2, S_3) \\
=&- \frac{1}{1+e^{-(s+t)+\tau}} + \frac{1}{1+e^{-s+\tau}} + \frac{1}{1+e^{-t+\tau}} \\
=&\frac{1}{\{1+e^{-(s+t)+\tau}\}\{1+e^{-s+\tau}\}\{1+e^{-t+\tau}\}} \\
&\times \{-(1+e^{-s+\tau})(1+e^{-t+\tau}) + (1+e^{-(s+t)+\tau})(1+e^{-t+\tau}) + (1+e^{-(s+t)+\tau})(1+e^{-s+\tau})\} \\
=&\frac{1}{\{1+e^{-(s+t)+\tau}\}\{1+e^{-s+\tau}\}\{1+e^{-t+\tau}\}} \\
&\times \{-1-e^{-s+\tau}-e^{-t+\tau}-e^{-s-t+2\tau}+1+e^{-(s+t)+\tau}+e^{-2t-s+2\tau}+e^{-t+\tau}+1+e^{-(s+t)+\tau}+e^{-2s-t+2\tau}+e^{-s+\tau}\} \\
=&\frac{1}{\{1+e^{-(s+t)+\tau}\}\{1+e^{-s+\tau}\}\{1+e^{-t+\tau}\}} \\
&\times \{-e^{-s-t+2\tau}+e^{-(s+t)+\tau}+e^{-2t-s+2\tau}+1+e^{-(s+t)+\tau}+e^{-2s-t+2\tau}\} \\
=&\frac{1}{\{1+e^{-(s+t)+\tau}\}\{1+e^{-s+\tau}\}\{1+e^{-t+\tau}\}} \\
&\times \{e^{2\tau}(-e^{-s-t}+e^{-2t-s}+e^{-2s-t})+2e^{\tau}e^{-s-t}+1\} \\
=&\frac{1}{\{1+e^{-(s+t)+\tau}\}\{1+e^{-s+\tau}\}\{1+e^{-t+\tau}\}} \\
&\times \{e^{2\tau}(-e^{-s-t}+e^{-2t-s}+e^{-2s-t}-e^{-2s-2t})+\{e^{\tau}e^{-s-t}+1\}^2\} \\
=&\frac{1}{\{1+e^{-(s+t)+\tau}\}\{1+e^{-s+\tau}\}\{1+e^{-t+\tau}\}} \\
&\times \{\{e^{\tau}e^{-s-t}+1\}^2+e^{2\tau}e^{-s-t}(-1+e^{-t}+e^{-s}-e^{-s-t})\} \\
=&\frac{1}{\{1+e^{-(s+t)+\tau}\}\{1+e^{-s+\tau}\}\{1+e^{-t+\tau}\}} \\
&\times \{\{e^{\tau}e^{-s-t}+1\}^2-e^{2\tau}e^{-s-t}(1-e^{-t})(1-e^{-s})\} \\
=&\frac{1}{\{1+e^{-(s+t)+\tau}\}\{1+e^{-s+\tau}\}\{1+e^{-t+\tau}\}} \\
&\times e^{2\tau-2s-2t}\{\{1+e^{-\tau+s+t}\}^2-(e^t-1)(e^s-1)\} \\
\geq&\frac{1}{\{1+e^{-(s+t)+\tau}\}\{1+e^{-s+\tau}\}\{1+e^{-t+\tau}\}} \\
&\times e^{2\tau-2s-2t}\{e^{-2\tau+2s+2t}-e^te^s\} \\
\geq&\frac{1}{\{1+e^{-(s+t)+\tau}\}\{1+e^{-s+\tau}\}\{1+e^{-t+\tau}\}} \\
&\times \{1-e^{2\tau-s-t}\} \\
\geq&0
\end{aligned}
$$

As a result, we have

$$D(S_1, S_3) \geq D(S_1, S_2) + D(S_1, S_3).$$

Thus, the triangle inequality holds if $\tau \leq \frac{s+t}{2}$.

Chapter 3

A complex network approach to study the extreme precipitation patterns in a river basin

Abstract

The quantification of spatial propagation of extreme precipitation events is vital in water resources planning and disaster mitigation. However, quantifying these extreme events has always been challenging as many traditional methods are insufficient to capture the nonlinear interrelationships between extreme event time series. Therefore, it is crucial to develop suitable methods for analyzing the dynamics of extreme events over a river basin with a diverse climate and complicated topography. Over the last decade, complex network analysis emerged as a powerful tool to study the intricate spatiotemporal relationship between many variables in a compact way. In this study, we employ two nonlinear similarity measures, namely, event synchronization and edit distance to investigate the extreme precipitation pattern in the Ganga river basin. We use the network degree to understand the spatial synchronization pattern of extreme rainfall and identify crucial sites in the river basin with respect to potential prediction skills. The study also attempts to quantify the influence of precipitation seasonality and topography on extreme events. The findings of the study reveal that (1) the network degree is decreased in the southwest to northwest direction, (2) the timing of 50th percentile precipitation within a year influences the spatial distribution of degree, (3) the timing is inversely related to elevation, and (4) the lower elevation greatly influences connectivity of the sites. The study highlights that edit distance could be a promising alternative to analyze event-like data by incorporating event time and amplitude when constructing complex networks of climate extremes.[1]

[1]Originally published as (P2): Ankit Agarwal, Ravi Kumar Guntu, **Abhirup Banerjee**, Mayuri Ashokrao Gadhawe, and Norbert Marwan: A complex network approach to study the extreme precipitation patterns in a river basin, Chaos 32, 013113 (2022) https://doi.org/10.1063/5.0072520

Extreme precipitation networks are constructed over a river basin using event synchronization (ES) and edit distance (ED). Edit distance is an alternative to event synchronization accounting for the sequences and the magnitude of events. Network-based measure degree is employed to understand the spatial synchronization pattern of the extreme precipitation in the Ganga river basin (GRB). The influence of the topography and rainfall characteristics on the extreme precipitation networks is also quantified. The proposed study can estimate the impact of orographic boundaries, thereby better understanding the extreme precipitation network's topology and spatial risk quantification of extreme events.

3.1 Introduction

Extreme events are of great societal interest as they may lead to many meteorological hazards that majorly affect lives and economic assets (Kalyan et al., 2021; Guntu and Agarwal, 2021). The evolution of such events in hydrology is non-uniformly distributed with time resulting in irregularly spaced data series. Traditionally, linear measures like Pearson and Spearman correlation are recognized as effective and robust estimators to quantify linear dependency between any two variables in irregular time series (Gadhawe et al., 2021). The Pearson correlation coefficient computes the linear dependency by looking into entire distribution of the data. For an extreme event-like data (tail part of the distribution), its application becomes suboptimal (Hlinka et al., 2014). Likewise, other linear approaches, such as wavelet analysis and Fourier transform are insufficient to capture nonlinear interrelationships (Marwan et al., 2007). As a result, defining a principled nonlinear approach is required to analyze extreme event time series, particularly when the dependency between event time series across locations is explored.

Numerous methods are available to study nonlinear dynamical behavior, such as recurrence quantification analysis, (Marwan et al., 2007; Banerjee et al., 2021; Donner et al., 2010), phase difference (Rosenblum et al., 1997; Schiff et al., 1996), Kolmogorov entropy (Pincus, 1995; Arnhold et al., 1999), or mutual information (MI) (Donges et al., 2009c; Runge et al., 2012), etc. However, by definition, extreme events are those that occur rarely and are separated by long periods. So, defining a suitable distance measure that can aid in analyzing the dynamics of extreme event-like time series becomes important in such instances. In the last decade, event synchronization (ES) has gained popularity (Quian Quiroga et al., 2002), particularly in climate science studies. The strengths of ES include capturing dynamic time-delays between various spatially distributed processes and thereby determining typical spatiotemporal patterns in climate systems and synchronization strength. Malik et al. (2012) used ES to analyze the extreme precipitation events (EPEs) in the Indian Monsoon System. Boers et al. (2013) conducted a similar study to characterize the extreme precipitation synchronicity in the South American Monsoon System. Later, Agarwal et al. (2017) proposed a multi-scale event synchronization method by combining the wavelet transform and ES. Furthermore, Kurths et al. (2019) used wavelet-based ES to unravel the spatial diversity of Indian rainfall teleconnections. An overview by Moreno and Perc (2019) complements nicely the hugely important role that network science has played in the analysis of various extreme events.

One can examine event-like data in its unaltered form by considering it as a discrete event series generated by a point process. As a measure of similarity, Victor and Purpura (1997) proposed a distance metric for calculating the distance between two spike trains (binary event sequences). Later, Hirata and Aihara (2009) developed a method for transforming a spike train into a time series known as edit distance. Then, this method was used by Suzuki et al. (2010) to construct a recurrence plot by measuring a distance between marked point processes to

analyze financial data. Later, several studies employed similar distance measures for analyzing irregularly sampled paleoclimate data (Eroglu et al., 2016; Ozken et al., 2015). Recently, Banerjee et al. (2021) proposed a modification of the edit distance (ED) to investigate dynamics of extreme event-like time series by integrating a nonlinear cost function. The robustness of the ED measure is demonstrated by several prototypical examples, and real flood event series constructed from discharge data of the Mississippi River in USA. However, to the best of the authors' knowledge, there is no study comparing the ES and ED-based extreme precipitation networks over a river basin with a diverse climate (Yadav et al., 2020) and with a complicated topography like the Ganga river basin (GRB) in India.

The GRB is highly vulnerable to EPEs due to its complex topography and altitude-dependent climate. Spatiotemporal changes in the EPEs at a basin-scale are crucial for preventing and mitigating water-related disasters and providing critical information for successful water resource management. EPEs over the Himalayan region are considered to be caused by a combination of thermodynamics and orographic uplifting (Thayyen et al., 2013). However, some studies argue that topography has no direct influence on precipitation (Barros et al., 2004). Conversely, Houze Jr (2012) identified topographic variables as fundamental contributors to precipitation fallout. Topography significantly impacts precipitation, but the relationship between topographical characteristics and precipitation is poorly understood (Palazzi et al., 2013; Wulf et al., 2010). Bharti et al. (2016) made an effort to study the EPEs in association with the elevation and found that the frequency and intensity of EPE exhibit an inverse relationship with elevation. Furthermore, basin-scale studies investigating the influence of topography on EPEs are rare. Therefore, there is a need to study the impact of topographic measures on the EPEs over the river basin having complex terrain in order to quantify the variability of EPEs.

In the present study, event synchronization and edit distance are compared to characterize the EPEs in the Ganga river basin; this involves constructing extreme precipitation networks over the basin. The topology of the resulting networks is quantified using graph theory-based measures such as network degree. Finally, a statistical comparison is made between the network-based measures and topographic measures to quantify the impact of topography on EPEs.

3.2 Study area and dataset

3.2.1 Study area

The Ganga river is the longest river in India and is located in the northern part of the country, covering the western and central Himalayas. The GRB is a transboundary basin and lies in India, China, Nepal, and Bangladesh. The present study is carried out over the region falling in India only due to gauge-based observations. The spatial extent of GRB falling in India covers 21°6′ N to 31°21′ N and 73°2′ E to 89°5′ E in longitude and latitude directions, respectively. An undulating terrain pattern characterizes the topography of GRB. The elevation ranges from about 7000 m north to less than 100 m above mean sea level in the south [Fig. 3.1(a)]. This complex topography leads to high spatiotemporal variability in precipitation and diverse climate classifications ranging from semi-tropical to semi-arctic (Bharti et al., 2016). The GRB's typical average annual rainfall ranges from 300mm at the western side to 3700 mm at the eastern end [Fig. 3.1(b)]. The GRB experiences extreme precipitation not only due to monsoon but also because of intense tropical cyclone activity pre- and post-monsoon. Flow in the basin is primarily driven by the precipitation coming from the Indian summer monsoon (June to September) and snowmelt from the Himalayas during the Spring season (March to May).

3.2.2 Dataset

In the present study, we use the gridded daily precipitation (mm) data with a high spatial resolution of $0.25° \times 0.25°$ covering the entire GRB. The gridded dataset was generated from a network of 6995 gauging stations across India using the inverse distance weighted interpolation technique proposed by Shepard (1968). The dataset was developed by the Indian Meteorological Department (IMD) (Pai et al., 2014) and can be retrieved from IMD's website. We use daily data for 22 years (1998-2019) to examine the EPEs over the GRB. In the past, several studies used the same dataset for various hydro-meteorological applications. These include extreme precipitation analysis (Rathinasamy et al., 2019), precipitation regionalization (Guntu et al., 2020a), intrinsic predictability of Indian summer monsoon (Guntu et al., 2020c), spatiotemporal variability of precipitation over India (Guntu et al., 2020b), and spatial diversity of Indian precipitation teleconnections (Kurths et al., 2019), among others.

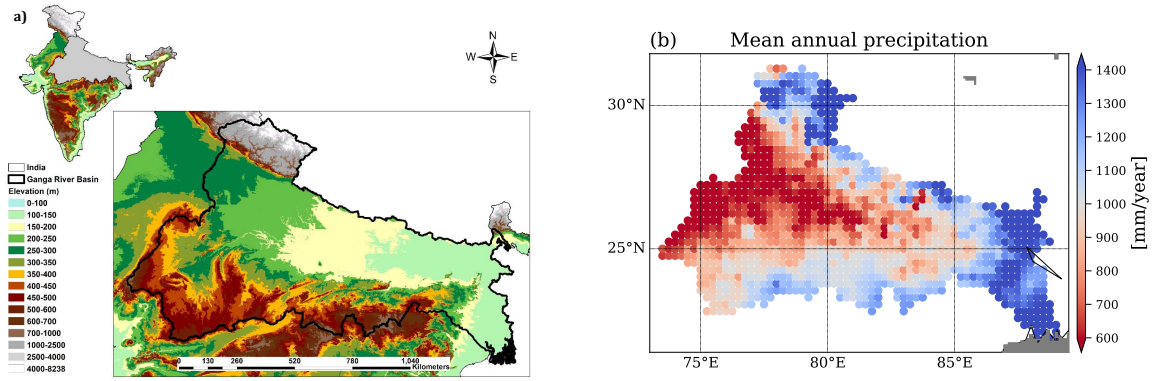


Figure 3.1: (a) Geographical location of the Ganga river basin in India showing elevation of the study area in meters above mean sea level. The elevation map is prepared by using the shuttle radar topography mission digital elevation model. This figure is generated using ArcGIS 10.6 (https://www.esri.com/en-us/arcgis/products/arcgis-pro). (b) Spatial distribution of mean annual precipitation (in mm) over the period 1998 -2019 based on IMD observations (3.2.2)

3.3 Methodology

In this work, we use two event-based nonlinear similarity measures, event synchronization (ES) (Quian Quiroga et al., 2002) and edit distance (ED) (Victor and Purpura, 1997), to study the extreme precipitation pattern across the GRB. In ES and ED, we consider events or sequences of events at one grid point and measure the strength of synchronicity with the other grid points. In this way, we construct the similarity matrix for all the grid points. This enables us to understand the underlying pattern of extreme events in more detail with respect to elevation and precipitation seasonality. In addition to the timing of events, ED also considers the similarity based on amplitude variations.

3.3.1 Edit distance

Victor and Purpura (1997) developed a specific distance metric to measure the similarity between the two spike trains based on the cost to transform one spike sequence into the other one. Hirata and Aihara (2009) extended this idea for converting a spike train into time series and named it edit distance. Edit distance was used for marked point processes to study the recurrence properties of exchange tick data of foreign currencies (Suzuki et al., 2010), recurrence analysis of irregularly sampled data (Ozken et al., 2015), and extreme event-like data (Banerjee et al., 2021).

We consider the event series at every grid point as a spike train. Let us consider two grid points i and j to find the similarity between the events; we transform one spike train into another by using the following operations: shifting in time, change in amplitude, and deletion/insertion [Fig. 3.2(a)]. Each of these operations is assigned with a cost. We use a combination of all three elementary operations and find the minimum path to convert one segment to another. The minimum transformation cost is defined as (Ozken et al., 2015, 2018)

$$D(S_i, S_j) = \sum_{(l,m)\in C} \{p_0\|t_l^i - t_m^j\| + p_1\|L_l^i - L_m^j\|\} + p_2(\mid I \mid + \mid J \mid -2 \mid C \mid) \tag{3.1}$$

where C is a set of pairs involved in the shifting process; the first term on the right-hand side of Eq. (3.1) corresponds to the shifting of an event by time $t_j - t_m$ and change in amplitude by $L_j - L_m$, and the second term involves the deletion/insertion of events. t_l^i and t_m^j denote the time of events and L_l^i and L_m^j denote the amplitude of events.

In edit distance analysis, we have three cost parameters, namely, p_0, p_1, and p_2, which are associated with the cost of shifting in time, change in amplitude, and cost of deletion/insertion, respectively. In the present study, we adopted the definition of the cost parameters proposed by Ozken et al. (2015, 2018). The first cost parameter p_0 provides the weightage for shifting in time for two event series and is calculated as $p_0 = \frac{M}{\text{total time}}$, where M is the number of events. So, when we calculate the transformation cost between two events series, we count the total number of events in two grid points divided by the total time, and it could be considered an event rate. The dimension of this parameter is 1/time. The second cost parameter is responsible for the change in amplitude,given by $p_1 = \frac{M-1}{\sum_1^{M-1}||x_i - x_{i+1}}||$ where x_i is the amplitude of i^{th} event in the data. The dimension of this parameter is 1/Length. The cost of deletion/insertion p_2 is set to be 1,

$$D(S_i, S_j) = \sum_{(l,m)\in C} \{p_0\|t_l^i - t_m^j\| + p_1\|L_l^i - L_m^j\|\} + (\mid I \mid + \mid J \mid -2 \mid C \mid) \tag{3.2}$$

We apply the distance measure Eq. (3.2) for all the pairs of grid points and store the value as a similarity matrix R_{ij}. In the present study, the cost of shifting in time p_0 depends on the number of events M. But, as we fixed the number of events in each grid point, the relative contribution in every pair of event series would be the same. Hence, the cost for shifting operation (in time + difference in amplitude) is mainly influenced by the parameter p_1. Unlike p_0, the value of p_1 will change depending on the difference in amount of rainfall. So, the final cost is mainly influenced by parameter p_1. ED works recursively, i.e., the shifting operation is associated with the events of more or less similar time and amplitude. So, in this way, ED can distinguish between the events with small variations in time and amplitude and with the events with a large difference in time and amplitude. This work chooses $\sim$ 36 events (top 10 percentile in a year) for each grid point and constructs the event series. The threshold value may differ from grid to grid. In this way, we manage to find a way to reduce the bias due to different event numbers.

3.3.2 Event synchronization

Event synchronization is a method to measure the synchronicity between climate extremes like extreme precipitation (Boers et al., 2019) or heatwave pattern (Mondal and Mishra, 2021). In precipitation network analysis, we consider each precipitation grid point as a node. To measure the synchronicity of events between two grid points i and j, let us consider an extreme event l occurring at a time t_l^i supposed to be synchronized with another event m at grid point j at time t_m^j with in a time lag $\pm T_{lm}^{ij}$ [Fig. 3.2(b)], where T_{lm}^{ij} is a flexible time lag that adopts

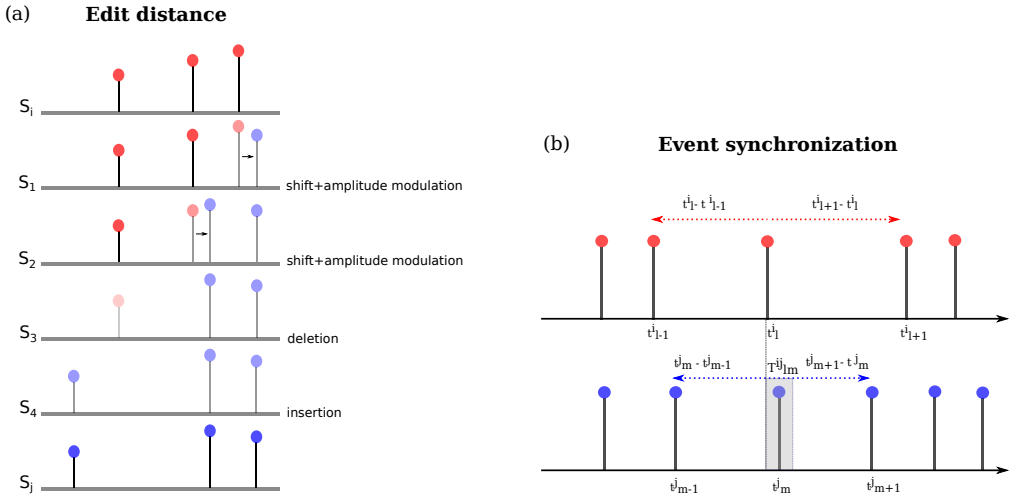


Figure 3.2: Schematic illustration of (a) the edit distance method: segment S_i transformed to S_j by following the path from S_1 to S_4 operations; the final cost is obtained by minimizing all the possible combination. (b) The event synchronization: t_l^i and t_m^i are occurrence of events in the event series at grid point i and j, and T_{lm}^{ij} is the adaptive time lag.

the density of events (sparse events=larger lags, dense events=smaller lags) and is expressed as follows:

$$T_{lm}^{ij} = \frac{min(t_{l+1}^i - t_l^i, t_l^i - t_{l-1}^i, t_{m+1}^j - t_m^j, t_m^j - t_{m-1}^j)}{2}. \tag{3.3}$$

Here, $l = 1, 2, 3, ..., S_i$ and $m = 1, 2, 3, ..., S_j$ are the number of events at i^{th} and j^{th} grid points, respectively. We count the coincidence of events at grid points i and j with the following scheme (Ozturk et al., 2018):

$$J_{ij} = \begin{cases} 1 & \text{if} \quad 0 < t_l^i - t_m^j < T_{lm}^{ij}, \\ \frac{1}{2} & \text{if} \quad t_l^i = t_m^j, \\ 0 & \text{else.} \end{cases} \tag{3.4}$$

The synchronization is measured by counting the weights of the synchronized events in both the time series $c(i|j) = \sum_{l=1}^{s_i} \sum_{m=1}^{s_j} J_{ij}$ and vice versa : $c(j|i)$ at the grid points i and j. Finally, the strength of synchronization is defined as (Malik et al., 2012)

$$Q_{ij} = \frac{c(i|j) + c(j|i)}{\sqrt{(s_i - 2)(s_j - 2)}} \tag{3.5}$$

where Q_{ij} lies in the range of $0 - 1$, where 1 implies the complete synchronization and 0 the absence of synchronization. We use this method for all pairs of grid points ($i \neq j$) and obtain the similarity matrix, which stores the strength of synchronization between extreme precipitation at all grid points.

3.3.3 Study design to test the efficacy of ED and ES

Finding a suitable similarity measure is important to investigate the interactions between different time series. However, when we deal with event-like data, the available options of similarity measures become very specific and have some limitations.

Kreuz et al. (2007) carried out a study to compare the synchronization measurements using different similarity measures. We adopted a similar prototype system to study the interaction

between two time series and their strength of synchronization for different coupling values and is described as follows. Consider two unidirectionally coupled Hénon maps with the following equations (Schiff et al., 1996):

$$x_1(i+1) = 1.4 - x_1^2(i) + bx_2(i);\ x_2(i+1) = x_1(i) \tag{3.6a}$$

$$y_1(i+1) = 1.4 - [\mu x_1(i)y_1(i) + (1-\mu)y_1^2(i) + by_2(i);\ y_2(i+1) = y_1(i) \tag{3.6b}$$

where x_1 and x_2 are the components of the drive system, y_1 and y_2 are the components of the response system, and μ is the coupling coefficient that varies in range of $0-0.8$. A similar model system has been adopted in previous studies (Paluš et al., 2001; Quian Quiroga et al., 2000; Romano et al., 2007) to investigate the direction of coupling. In the first step, the event series is constructed. To do so, we scan through x_1 and y_1 component and find the local maxima $(x(t_i) > x(t_i \pm 1))$ and store the times t_i^x, t_j^y, and amplitudes L_i^x, L_j^y $(i = 1, 2, ..., n_x;\quad j = 1, 2, ..., n_y)$, where n_x and n_y denote the number of events. The local maxima procedure has been successfully used to study event-like data (Kreuz et al., 2007; Quian Quiroga et al., 2000; Odenweller and Donner, 2020).

When the coupling is small [see x axis in Fig. 3.3], the similarity between two time series is less; hence the Q value is low and, D value is high as the cost of transformation is very high. As we increase the coupling parameter (μ), the value of D decreases [Fig. 3.3(a)], and the value of Q increases [Fig. 3.3(b)]. We detected the first four transition points using change-point detection (Killick et al., 2012), where abrupt changes in Q and D values are detected. In the previous studies (Paluš et al., 2001; Quian Quiroga et al., 2000; Romano et al., 2007), the onset of identical synchronization was found to start for $\mu \geq 0.6$. To confirm the direction of interdependency, a mutual information (MI) estimator is used between two time series x_1 and y_1. The original time series is considered in the analysis but not the local maxima from the driver and response system. Kraskov et al. (2004) proposed a method to compute mutual information based on the k-nearest neighbor. MI measure has less bias and is robust against short time series (Vejmelka and Paluš, 2008). A detailed description of MI method is not provided here, for detail information of MI, readers are requested to refer to the study by Kraskov et al. (2004) for detailed methodology.

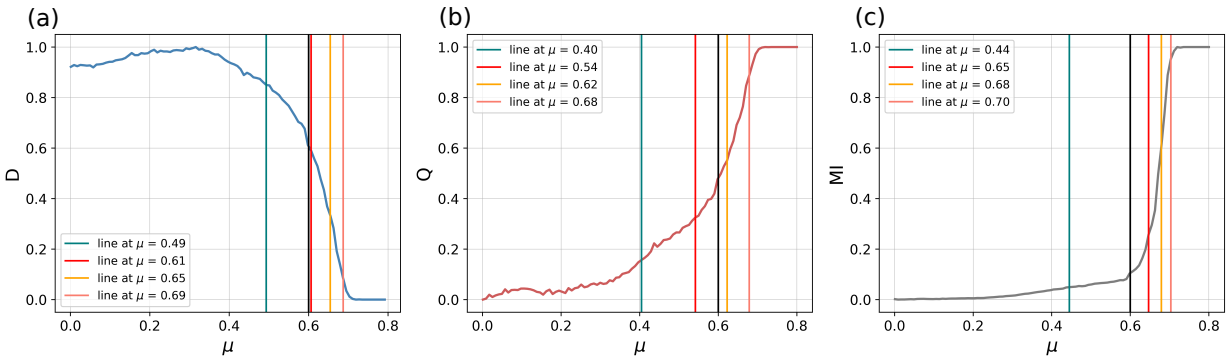


Figure 3.3: Detection of changes in the direction of coupling with coupling parameter value (μ) ranging from 0 to 0.8 for (a) event synchronization (ES) and (b) edit distance (ED) with (c) mutual information (MI). Vertical color lines represent the transition in the direction of coupling. The black vertical line signifies the onset of identical synchronization

Interestingly, from Fig. 3.3, we observed that all transition points lie early for relatively low coupling in ES [Fig. 3.3(b)] when compared with the mutual information [Fig. 3.3(c)]. However, in the case of ED [Fig. 3.3(a)], the transition points (vertical colored lines) are placed when the coupling is relatively strong, i.e., the similarity between two time series is high. The prototypical system shows that both ED and ES are suitable enough to study the

interdependence between two event-like time series. Event series with almost coinciding events but strong differences in amplitudes will be considered very similar by ES but more different by ED; only the event with similar occurrence and amplitude variation can be considered more similar by ED. With the additional information of amplitude incorporated in ED, it helps in capturing a more realistic picture

3.3.4 Network construction and network measure

We obtain the adjacency matrix by applying a certain threshold upon the similarity matrix. In this study, the threshold is defined based on the link density [Eq. (3.7)],

$$\rho = \frac{2E}{N(N-1)} \tag{3.7}$$

where E is the number of edges and N is the number of nodes, and the corresponding threshold $\tau = \tau(\rho)$ is chosen. For ES, the adjacency matrix defined as

$$A_{ij}^{ES} = \begin{cases} 1 & \text{if} \quad Q_{ij} > \tau, \\ 0 & \text{else.} \end{cases} \tag{3.8}$$

For ED, the adjacency matrix is defined as

$$A_{ij}^{ED} = \begin{cases} 1 & \text{if} \quad R_{ij} < \tau, \\ 0 & \text{else.} \end{cases} \tag{3.9}$$

For both ED and ES, we choose the threshold τ for an edge density of 0.05. In the previous climate network applications (Malik et al., 2012; Wolf et al., 2021; Stolbova et al., 2014), a link density of 0.05 satisfies a high similarity and retains the significant links.

In the study of complex networks, there are multiple network measures available based on the research question. Our research is particularly interested in finding the grid points that show a similar pattern of extreme events. Accordingly, the node degree (k) of a node i is calculated, indicating the number of connections with other nodes in a network,

$$k_i = \sum_{j=1}^{N} A_{ij}, \tag{3.10}$$

where N is the number of nodes and A_{ij} is the adjacency matrix.

3.3.5 Precipitation seasonality

In the following, we assess the difference in node degree (k) obtained from event synchronization and edit distance, which can be applied in understanding the difference in the topology of the extreme precipitation network patterns. Feng et al. (2013) calculated the first moment of area for monthly precipitation to derive the timing of 50^{th} percentile precipitation within a year [Fig. 3.4(a)]. Later, several studies (Guntu et al., 2020a; Pascale et al., 2015) used the same centroid concept to derive the timing (T_i) of 50^{th} percentile precipitation over different time scales.

The calculation of the long-term average timing is a two-step procedure. We considered 22 years of data, and the median value of precipitation is estimated for all the months [Fig. 3.4(a)]. Next, the timing (T_i) was computed using Eq. (3.11) and is indicated in Fig. 3.4(b),

$$T_i = \frac{\sum_{m=1}^{12} mP_m}{\sum_{m=1}^{12} P_m} \tag{3.11}$$

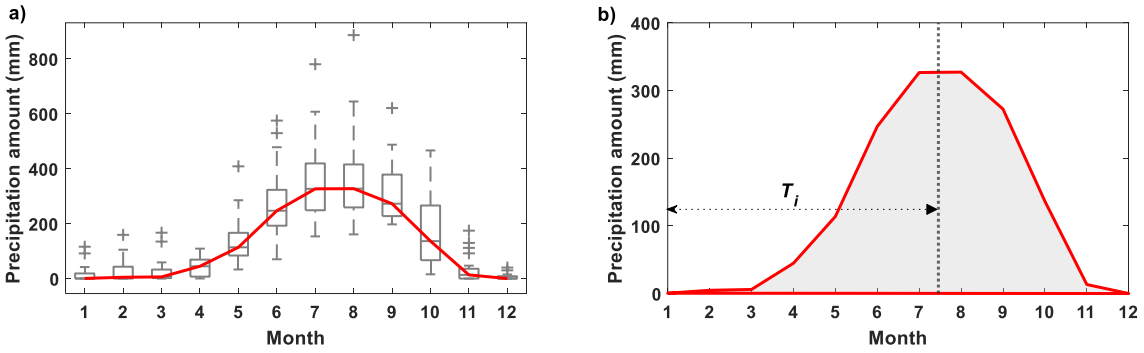


Figure 3.4: (a) Long-term average monthly precipitation time series for one grid point (red color). The box-plot statistics of each month are represented in gray color. (b) Computation of the timing of 50^{th} percentile precipitation (T_i) linked with the area of monthly precipitation. The vertical dotted line represents the timing of the peak precipitation.

where m represents the month number from 1 to 12 and P_m represents the corresponding precipitation amount during that particular month.

3.4 Result and discussion

We present the results in four categories: first, investigating network characteristics of extreme precipitation; second, linking network characteristics with precipitation seasonality; third, with topographic variability; and finally, quantifying node importance over GRB.

3.4.1 Network characteristics of extreme precipitation over the Ganga river basin

Network characteristics are obtained by thresholding the similarity matrix Q_{ij} for ES and R_{ij} for ED. Network characteristics depend on the choice of threshold τ. In the present study, we constrain the link density $\rho = \frac{2E}{N(N-1)}$ where E gives the total number of edges in the current network and the denominator is the number of all possible edges in the network (if fully connected), with the number of nodes N and selecting the corresponding threshold $\tau = \tau(\rho)$ (Donges et al., 2009c). In both cases, we choose the threshold so that it corresponds to 5% link density. The precipitation network comprises nodes representing the time series of each grid point and the edges showing statistically significant interdependencies between the nodes (Tsonis et al., 2006).

We compute the node degree for all the grid points in the GRB, following the procedure described in Sec. 3.3.4. Figures 3.5(a) and 3.5(b) present the spatial distribution of the degree of each node using edit distance and event synchronization. We observe significant variations in the spatial distribution of node degree values across the GRB. The grid points with high degree (more links) are more likely to be found in the southwest part of the basin, while degree values decrease toward the northeast part and periphery of the basin in both ED- and ES-based networks. In general, a high degree of a node (higher connectivity) indicates that the extreme precipitation data at that grid point shares a piece of similar information with many other grid points over the basin and is likely to have greater impact on functioning of the network. Since similar information is shared, it is not considered a unique one. On the other side, the grid point with degree zero does not share similar information with any other grid point.

The artificially imposed boundary (in this case, the Ganga river basin border) may impact the estimated network measures because spatial boundaries cut links that would connect the considered region with outside regions. We use the boundary correction procedure suggested

by Rheinwalt et al. (2012) to avoid spurious bias on node degree arising from this boundary effect. We use the procedure as follows; we construct 100 spatially embedded random networks (SERNs) that preserve the position of the nodes in space and link length probability depending on the spatial link length of the original network. We compute the network measure for all the SERN surrogates. The boundary effect is measured by taking the average of the network measure for all the surrogates. The corrected network measure is obtained by dividing the original measure with the estimated boundary error for all the nodes. The corrected network measure is dimensionless as it provides the network measure relative to the expected value of a SERN.

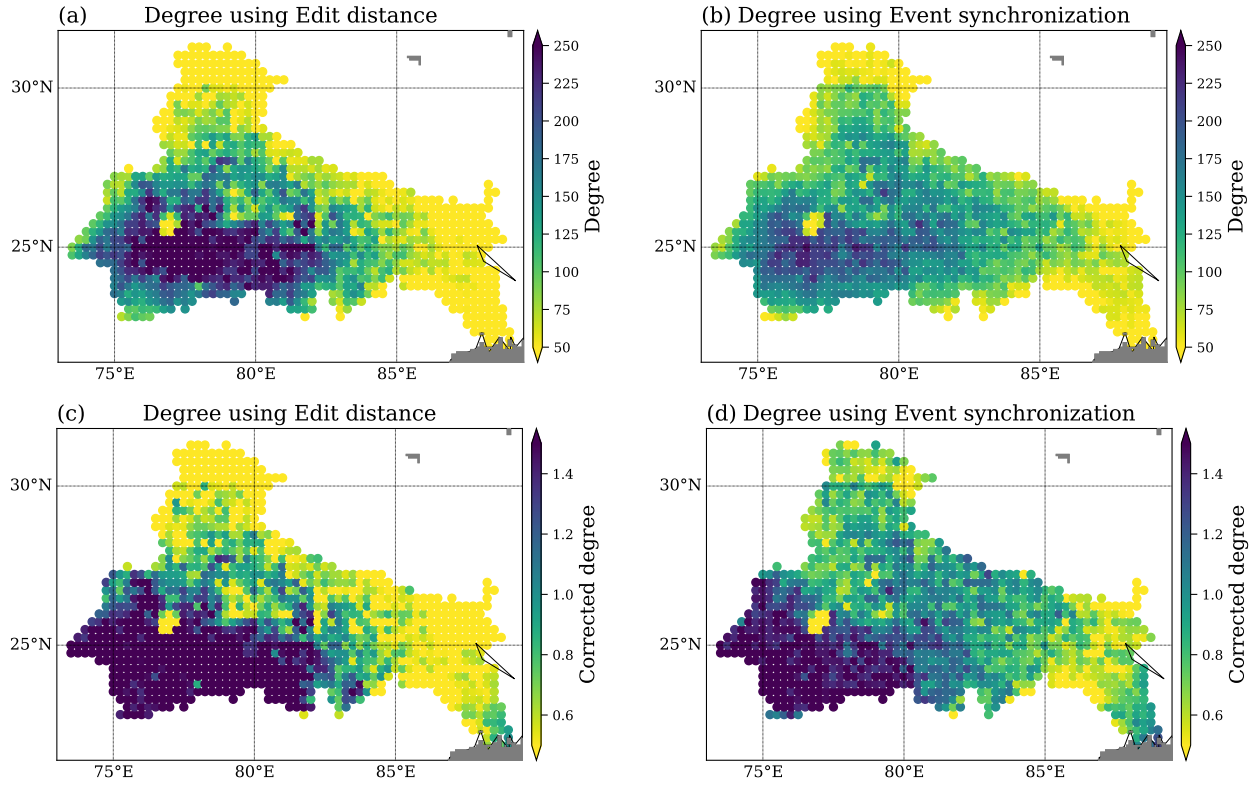


Figure 3.5: Spatial distribution of node's degree for (a) the edit distance and (b) event synchronization, and corrected degree for (c) the edit distance and (d) event synchronization.

Figures 3.5(c) and 3.5(d) present the spatial distribution of the corrected degree value computed by edit distance and event synchronization of each grid cell in the basin. In the case of ED, we find high degree concentrated at a particular (south- west) part of the basin. In contrast, the corrected degree value of the ES-based network is varying all over the basin, but a high number of nodes having high degree values are at the southwest part of the basin.

By comparing the results of the ES and the ED networks, we find a slight difference in the spatial distribution of degree values. The southwest part shows a high degree in the ED network [Fig. 3.5(c)] and spreads out to the bay's periphery. In the ES network [Fig. 3.5(d)], the degree value follows a similar pattern as ED networks but with lower degree values in the stronger connected region and with a larger extension of this stronger connected region.

Therefore, by comparing Figs. 3.5(a)-3.5(d), we see that these nonlinear methods efficiently capture the general network topology without being affected by the artificial boundary.

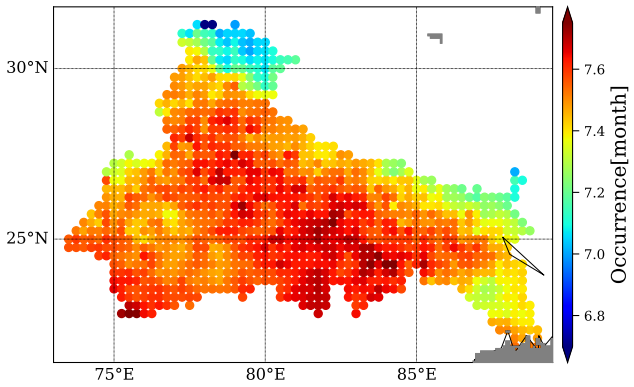


Figure 3.6: Spatial pattern of the timing of 50th percentile precipitation for the temporal window 1998-2019.

3.4.2 Relationship between network characteristics and precipitation seasonality

We quantify the centroid of the monthly precipitation distribution over the GRB to derive the timing of 50th percentile precipitation (see Sec. 3.3.5). The timing is mainly during June to August over the GRB (Fig. 3.6). The Himalayan region (north side) receives its 50th percentile precipitation at the beginning of June, and in contrast, in the middle of July over the plain region (south region) and at the end of July over the northeast and deltaic region (southwest region) of the GRB. To know if there is any statistical relationship between 50th percentile precipitation and elevation, we consider the occurrence of rainfall and elevation (Fig. 3.7).

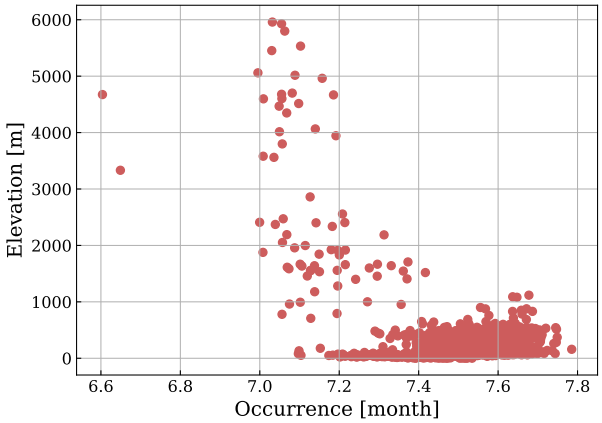


Figure 3.7: Relationship between the timing of 50th percentile precipitation and elevation.

Next, we consider the relationship between the timing of 50th percentile precipitation and the corrected node degree values for ED and ES (Fig. 3.8). In both cases, grid stations with early timing (before July) and intermediate timing (from July to mid-July) have lower degree values, i.e., fewer connections, while stations with relatively later timing have greater degrees varying from 0 to 2.7. For ED, the relationship between the timing and degree is more apparent than for ES due to the inclusion of the amplitude variations. Thus, by comparing Figs. 3.8(a) and 3.8(b), ED could be a promising alternative to analyze event-like data by incorporating event time and amplitude.

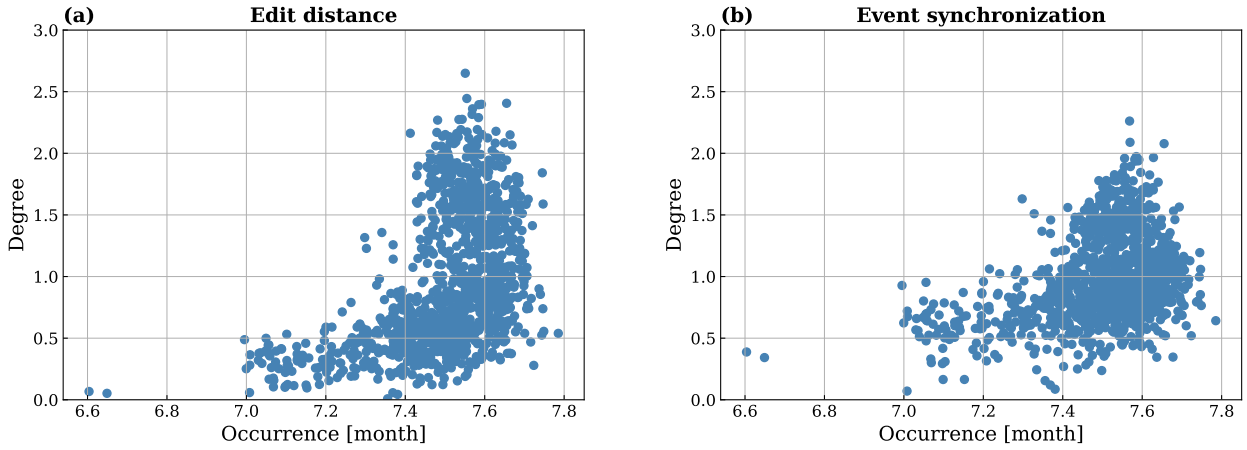


Figure 3.8: Relationship between occurrences of peak precipitation and the connectedness between the grid points (degree) for the network constructed using (a) edit distance and (b) event synchronization.

3.4.3 Relationship between network characteristics and topographic measure

We further consider the relationship between node degree and elevation of each grid cell to quantify the influence of elevation on the network topology (Fig. 3.9). The degree varies for the grid cells at different elevations. For example, sites with low elevation (<1000 m) have high degree values ranging from 0 to 2.5 and from 0 to 2.1 in ED and ES, respectively, while the stations at intermediate to high elevation tend to have fewer connections, i.e., low degree values (below 0.5 for ED and between 0.5 and 1 for ES). The degree in ED-based networks [Fig. 3.9(a)] is relatively lower than for ES-based networks [Fig. 3.9(b)] for stations located at elevations between 1000 and 6000 m.

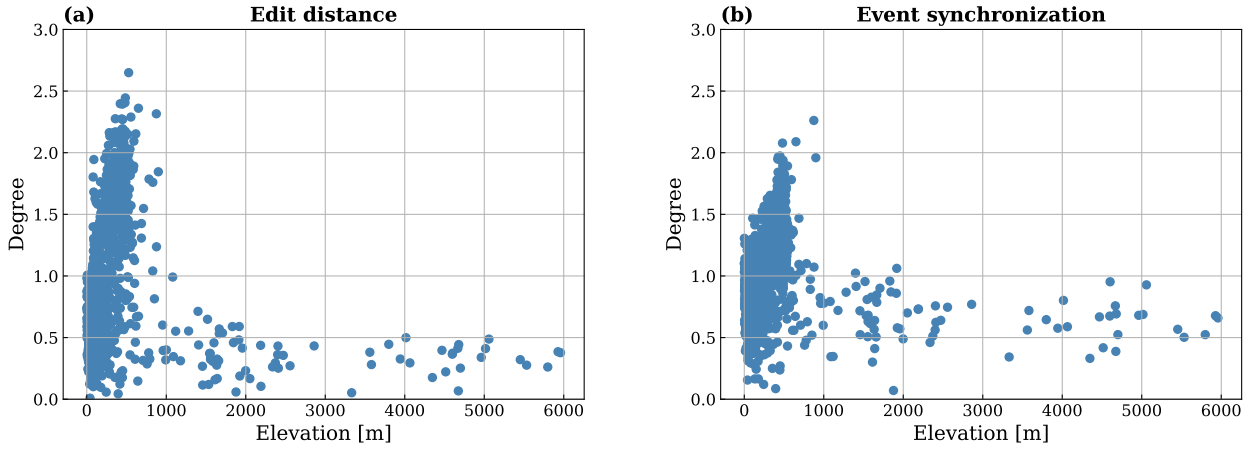


Figure 3.9: Relationship between degree and elevation for the precipitation network constructed using (a) ED and (b) ES.

The southwest of the GRB has moderate elevation and grid points over that region receive precipitation during the Indian summer monsoon only [point A in Figs. 3.10(a) and 3.10(b)]. The north of the GRB has the highest elevation compared to the rest of the basin [Fig. 3.1(a)] and is mainly dominated by western disturbances (Raj et al., 2021) during December to May and receives its highest precipitation during the Indian summer monsoon [point B in Figs. 3.10(a) and 3.10(b)]. As a result, the timing of 50^{th} percentile precipitation of the southwest region is more compared to the northern region. The degree and timing are low and the connectedness is limited to a few grid points only in the north of the GRB [Figs. 3.10(c) and 3.10(d)]. In

particular, connections made by ED [Fig. 3.10(c)] provide additional information as it includes amplitude as well. We find that grid points in the southwest have a higher degree but not to the central part of the GRB. The western and southwestern region receives less precipitation and has lesser spatiotemporal variability (Guntu et al., 2020b). As most of the grid points in the southwest region receive similar information, it has higher connectedness [Figs. 3.10(c) and 3.10(d)]. Interestingly, the timing is the same across the central part and southwest of the GRB (Fig. 3.6). The primary factor in getting a higher degree across the southwest region is the elevation. As a result, ED connections mostly lie in the same topographical region [Fig. 3.10(c)], whereas ES links are going to another topographical region as well [Fig. 3.10(d)].

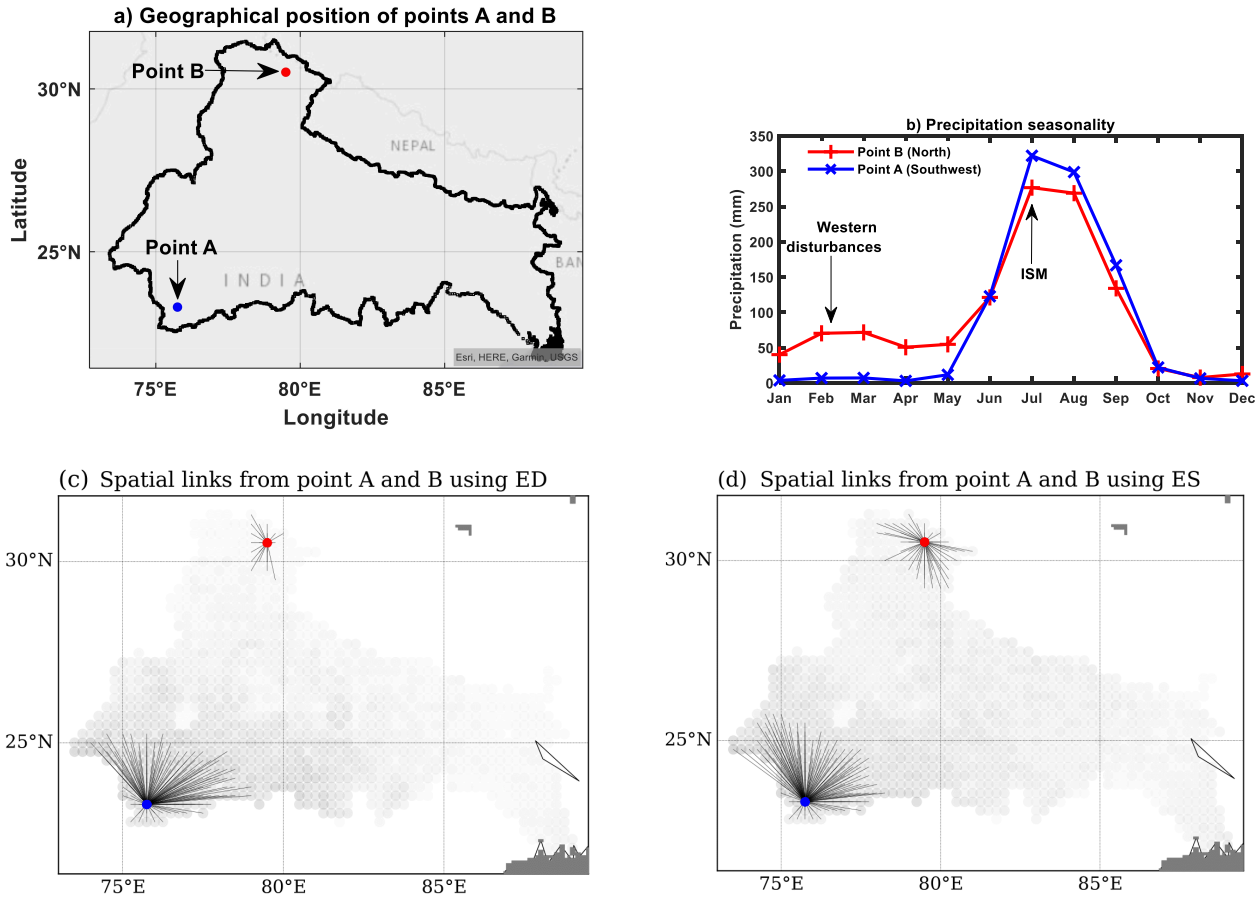


Figure 3.10: (a) Map of GRB showing points in the southwest and north and (b) line plot illustrating precipitation seasonality and their connections found using ED (c) and ES (d).

Therefore, by comparing Figs. 3.10(c) and 3.10(d), ED as a similarity measure provides additional information in segregating the extreme events originating during western disturbances and Indian summer monsoon.

3.4.4 Quantifying node importance over GRB

In several studies of complex network dynamics and structure, node importance is becoming a hot topic (Gadhawe et al., 2021; Agarwal et al., 2020). Node importance is evaluated and primarily researched regarding the network's structural properties and application requirements. It is critical to identify significant nodes to create effective networks. The essential nodes in the network are a small number of particular nodes that have a considerable impact on the network's structure and function (Agarwal et al., 2018). Therefore, we argue that those nodes of large importance might indicate locations that are suitable for representing the overall precipitation state/regime or for predicting the extreme events. Degree, closeness, betweenness,

clustering coefficient, weighted degree betweenness, and other centrality measures have been developed to evaluate node importance in complex networks (Agarwal et al., 2020). For a comparative understanding of node importance derived from the network constructed using ED and ES, corrected degree values obtained over GRB are used.

Figure 3.11 shows the spatial distribution of location and network connections of the top 5% of the highest degree grid cells in the ED [Fig. 3.11(a)] and ES-based network [Fig. 3.11(b)]. The 5% top degree nodes infer that the time series at the grid stations in the south and southwest region of the basin have more robust connectivity than the rest of the stations in the GRB. Compared to ED, in ES-based networks, important grid points are concentrated in the outer southwest region of the basin [Fig. 3.11(b)]. Furthermore, we plotted the connections corresponding to this high degree grid station to reduce the complexity in visualizing the interactions. Theoretically and practically, identifying node importance in complex networks is vital for enhancing network robustness and invulnerability.

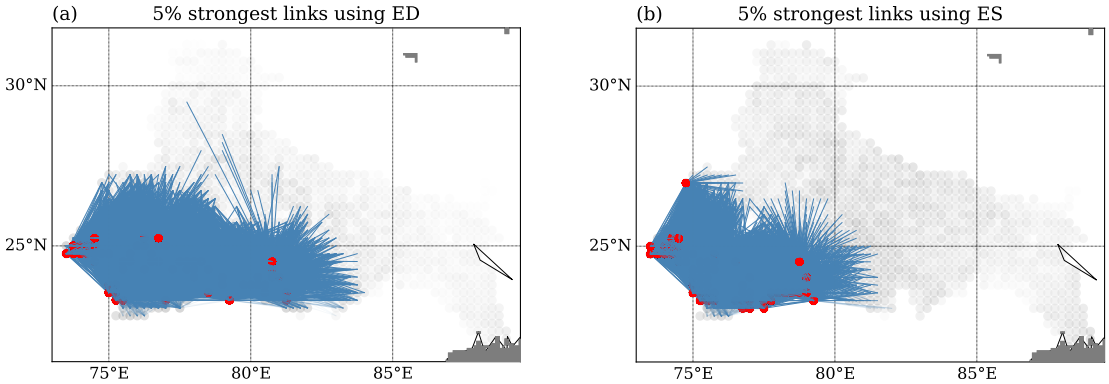


Figure 3.11: Spatial distribution of the strongest 5% of nodes obtained over the GRB using (a) ED and (b) ES. All 1174 grid stations are plotted in the background in light gray color. Red colored nodes indicate the top 5% degree nodes

3.5 Conclusion

In this work, daily precipitation records of 1174 grid points during 1998-2019 in the Ganga river basin are studied by their extreme precipitation network characteristics. Two nonlinear methods, event synchronization and the edit distance method are used to compare the extreme precipitation patterns. We compare the network topology resulting from the edit distance-based network ,and event synchronization-based network. The network measure node degree found that the degree is decreasing from the southwest to the northwest direction. As it is known that the extreme precipitation patterns are influenced by complicated topography, we quantify the variability of EPEs by investigating the impact of topography on the EPEs across the basin. We find that the connectivity of events is greatly affected by elevation. Our analysis revealed an inverse relationship between elevation and the timing of 50th percentile precipitation. The network-based degree suggests that the event-like time series at the grid stations located at the southwest region of the basin are more connected compared to the rest of the basin. When an extreme event occurs across sub-basins under this area, it contributes to a higher flooding risk.

Furthermore, as a similarity measure, edit distance includes amplitude variations in the pairwise tests. In our work, the degree pattern using ED indicates the node of high similarity in terms of their amplitude and occurrence. Precipitation extreme events with almost coinciding events but significant differences in amplitudes will be considered very similar by event synchronization but more different by edit distance. More minor synchronous extreme events

with very similar amplitude variation can, in contrast, be considered different by event synchronization but more similar by edit distance. One of the primary purposes of this research is to understand the spatial distribution of EPEs in a specific area. Although we illustrated the preliminary results about the spatial connectivity of EPEs, there is still a need to understand the propagation of EPEs at different spatial and temporal scales. Also, a statistical significance testing is required for more credible network interpretation. We recognize that the application of network theory in hydrological systems is still in its infancy. We believe that current research on applying complex networks to analyze the extreme precipitation pattern will improve our understanding of the spread of extreme events in the GRB.

Data availability

The dataset is developed by the Indian Meteorological Department (IMD) and can be retrieved from the IMD website (last accessed 1 September 2021), https://imdpune.gov.in/Clim_Pred_LRF_New/Grided_Data_Download.html.

Chapter 4

Spatial coherence patterns of extreme winter precipitation in the United States

Abstract

Extreme precipitation events have a significant impact on life and property. The United States experiences huge economic losses due to severe floods caused by extreme precipitation. With the varied terrain, it becomes increasingly important to understand the spatial variability of extreme precipitation to conduct a proper risk assessment of natural hazards such as floods. In this work, we use a complex network based approach to identify distinct regions exhibiting spatially coherent precipitation patterns due to various underlying climate mechanisms. To quantify interactions between event series of different locations, we use a nonlinear similarity measure, called the edit-distance method, which considers not only the occurrence of the extreme events but also their intensity while measuring similarity. Using network measures, namely, degree and betweenness centrality, we are able to identify the specific regions affected by landfall of atmospheric rivers in addition to those regions where the extreme precipitation due to storm track activity is modulated by different mountain ranges such as the Rockies and the Appalachians. Our approach provides a comprehensive picture of the spatial patterns of extreme winter precipitation in the United States due to various climate processes despite its vast complex topography. [1]

[1]Submitted to Theoretical and Applied Climatology (under review) as (P3): Abhirup Banerjee, Matthias Kemter, Bedartha Goswami et al. Spatial Coherence Patterns of Extreme Winter Precipitation In The United States, 17 January 2022, preprint available at Research Square doi.org/10.21203/rs.3.rs-1243541/v1

4.1 Introduction

Extreme precipitation poses a serious threat to lives and livelihood of people all around the world. With the intensification of extreme precipitation and flood events over most climate regions (Tabari, 2020; Easterling et al., 2017; Janssen et al., 2014; Vu and Mishra, 2019; Kunkel et al., 2012) due to climate change, understanding the spatial variability of extreme precipitation is crucial to manage the big socioeconomic losses often associated with them (Merz et al., 2021). Previous studies have shown that extreme precipitation connectivity in the US is highest during the winter months (Touma et al., 2018), while river flood connectivity is higher in spring in the Rocky mountains and central US due to snow melt (Brunner et al., 2020). As reported in the billion-dollar weather and climate disasters catalog released by the NOAA/National Centers for Environmental Information (NCEI), in the period 2010-2020, 328 people were killed due to flooding and winter storms in the U.S., and more than $77 billion (U.S. dollars) worth of economic damages were caused (Weather and for Environmental Information , NCEI). This included, for instance, the above-average precipitation leading to severe flooding in the Mississippi and Missouri Rivers and their tributaries during the winter season of 2019 (December 2018 – February 2019) (Hoell et al., 2021; Flanagan et al., 2020). Thus our study focuses on extreme precipitation in the winter months (DJF), during which rainfall can cause floods directly and snowfall leads to the accumulation of snow packs for the melting season.

Numerous studies have analyzed extreme precipitation events across the United States (Mondal et al., 2020; Najibi et al., 2020), meteorological causes of secular variations (Kunkel et al., 2012), spatiotemporal variability (Kursinski and Mullen, 2008), and their relation to large-scale meteorological patterns (Agel et al., 2019). Here, we focus on the spatial connectivity of extreme precipitation events, which is relevant for river flood generation and the spatial extent of simultaneous flooding (Brunner et al., 2020; Kemter et al., 2020). Understanding the spatial dependence of extreme precipitation and its underlying mechanism is important to assess risk from natural hazards. Simultaneous extreme precipitation across large scales can lead to synchronous flooding in multiple states, which has a greater societal and financial impact than independent, localized flood events due to regional interdependencies in risk management, infrastructure, and insurance (Jongman et al., 2014).

We use a complex network based approach to study spatial patterns of extreme winter precipitation in the United States. Climate network analysis can help to identify the regions which are most likely to experience concurrent precipitation extremes and which climatic conditions are responsible for their generation. The climate network belongs to the category of *functional network*, i.e., in which pairwise dependency of station or grid data is computed, and the network topology is represented by different network measures (Donges et al., 2009a; Fan et al., 2021b; Tsonis and Roebber, 2004). The network representation of climate data allows us to study pairwise interactions between climate variables of different locations. However, standard similarity measures, such as Pearson or Spearman correlation coefficients are not suitable to evaluate the relationship within extreme precipitation data which are event-like time series. In the past decade, a specific and nonlinear synchronization measure, in particular *event synchronization* (ES) (Quian Quiroga et al., 2002) have been used extensively to construct climate networks for event-like data such as extreme precipitation (Malik et al., 2011; Stolbova et al., 2014; Ozturk et al., 2019), heat wave pattern (Mondal and Mishra, 2021) etc. Boers et al. (2013, 2014c,a) used complex networks constructed based on ES to study the South American Monsoon and reveal the global extreme precipitation pattern (Boers et al., 2019). Konapala and Mishra (2017) used the same climate network framework to study hydroclimatic extreme events. Agarwal et al. (2017) introduced multi-scale event synchronization by combining wavelet transform and ES.

However, ES only considers the time of occurrence of events to identify the events coincidence and use this as a measure for similarity, but not the difference in strength or amplitude of the events. While very few previous works (Ciemer et al., 2018) have proposed some modified correlation measures to investigate spatial covariability pattern of general precipitation (also considering the amplitude variability), these methods are linear and, thus, not suitable to study extreme precipitation behaviour.

In our study, we use a special distance metric, particularly designed to study the similarity between spike trains, called *edit distance* (ED), first proposed by Victor and Purpura (1997) and later extended by Hirata and Aihara (2009). This metric has been used in combination with recurrence plot to analyze the recurrence property of marked point process data (Suzuki et al., 2010), paleoclimate data (Ozken et al., 2015, 2018), and extreme event-like hydrological data (Banerjee et al., 2021). Recently, Agarwal et al., (accepted) also applied this measure to study the extreme rainfall pattern in the Ganga River basin. Under this framework, we consider here each event series as a marked point process and measure the similarity between two such event series by optimizing the cost of transformation associated with transforming one event series to another one through elementary operations, such as shifting, addition or deletion of events.

Spatial patterns of different network measures, namely degree and betweenness centrality are used to study the spatial connectivity of extreme winter (December–January–February (DJF)) precipitation events. While the degree field is based on local topological information, the path-based betweenness centrality field includes global topological information (Donges et al., 2009a). Through our approach, we are not only able to identify regions with distinct extreme precipitation patterns, but also delineate the regions affected by atmospheric rivers and tornadoes. The patterns of Sec. 4.2, we describe in detail, the data and the methodology. In Sec. 4.3, we draw an interpretation based on our network analysis and discuss the results from a climatological point of view.

4.2 Data and Methodology

4.2.1 Data source and data pre-processing

In this study, we use daily averaged precipitation, geopotential height, wind at different pressure levels, and vertically integrated water vapour (IVT) flux data derived from ERA5 reanalysis (Hersbach et al., 2020) for the period 1980 – 2020. The spatial resolution used is $0.5° \times 0.5°$. It is worth mentioning here that although reanalysis precipitation data do show biases compared to the observations, observational datasets typically either have a limited spatial coverage (GPCC, TRMM, etc.), lower resolution (GPCP) or a limited temporal coverage (TRMM). The ERA5 shows in most cases, smaller biases than other reanalysis datasets (JRA-55, MERAA-2)(Hassler and Lauer, 2021). However, we verify the robustness of our results by comparing them with those obtained using JRA-55 (Japan Meteorological Agency, Japan, 2013) (see figures in the Supporting information).

4.2.2 Network construction

A network or graph comprises two main components: a set of nodes V and a collection of edges E. Mathematically, a network is expressed as $G = \{V, E\}$ (Sivakumar and Woldemeskel, 2014; Donges et al., 2009a). In case of climate network, each geographical grid point of the climate dataset is considered as a node and an edge is placed when there is statistically significant association or functional dependency between two nodes. To construct the climate network for extreme precipitation, first we transform the precipitation time series data at each grid

point into an extreme precipitation event series by considering those days as events for which precipitation is among the highest 5% of all values, including dry days without precipitation, in a particular season at that location, resulting in 4 to 5 events for each season (Malik et al., 2011; Boers et al., 2013; Stolbova et al., 2014). Here, we construct the network for extreme precipitation events to study its pattern of spatial variability.

In this study, we use the edit distance (ED) method, which takes into account both the sequence and amplitude of events. In general, ED is a distance measure to quantify the similarity/ dissimilarity between two spike trains (Victor and Purpura, 1997; Banerjee et al., 2021). Additionally, ED considers each event series as a marked point process (Suzuki et al., 2010; Ozken et al., 2015, 2018). The idea is to transform an event series into another series by performing some elementary operations: shifting in time, amplitude modulation, and deletion/insertion of events (Fig. 4.1). A specific cost is assigned to each operation. The total cost of transformation from one event series to the other is computed by tracing the minimal-cost path.

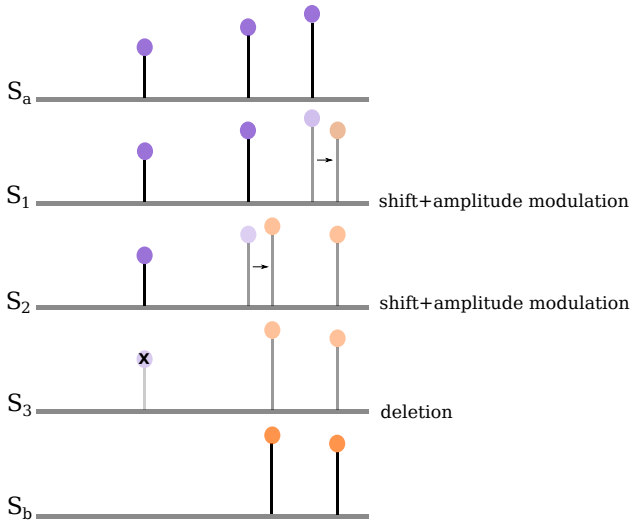


Figure 4.1: Schematic of the transformation of segment S_a to S_b through four steps numbered as steps $S_1, \ldots, S_3$. The path shown is a minimal-cost path and all steps are elementary steps, i.e., shifting an event, amplitude modulation, deleting/inserting.

The mathematical formulation of the distance measure is described as follows. Consider two given segments S_a and S_b, the minimum cost of transformation is defined as

$$D(S_a, S_b) = \min_{C} \Big\{ \Big\{ \sum_{(\alpha,\beta)\in C} \Lambda_0 \|t_a(\alpha) - t_b(\beta)\| + \Lambda_1 \|L_a(\alpha) - L_b(\beta)\| \Big\} + \Lambda_s(\mid I \mid + \mid J \mid -2 \mid C \mid) \Big\}. \tag{4.1}$$

The time and amplitude of events are denoted as $t_a(\alpha)$, $t_b(\beta)$, and $L_a(\alpha)$, $L_b(\beta)$, Λ_0 and Λ_1 are the coefficient of cost of shifting in time and amplitude change. The first term of Eq. (4.1) sums the cost of shifting in time and amplitude change between the α^{th} event in S_a and β^{th} event in S_b, C contains all the pairs associated in this operation. The second term in Eq. (4.1) denotes the deletion/ insertion operation, $\mid I \mid$, $\mid J \mid$ are the sets indices of events in S_a and S_b, Λ_s is the coefficient of cost of deletion.

The minimum cost implies the highest similarity and vice-versa. We then calculate the transformation cost for every pair of event series i and j of the gridded extreme event dataset using the above method, which gives us the similarity matrix Q_{ij} (here, cost matrix). Thereafter,

we obtain the adjacency matrix A_{ij} by thresholding the similarity matrix Q_{ij} with a suitable threshold, which gives the edges of our network. Mathematically, $A_{ij} = \Theta(\epsilon - Q_{ij}) - \delta_{ij}$, where Θ is the Heaviside function, i.e., we assign 1 when the cost is below a certain threshold, otherwise 0. ϵ is the threshold, and δ_{ij} is the Kronecker delta to remove self loops. In the case of ED, lower transformation cost between two event series implies higher similarity. For all pairs of grid cells whose value of the transformation cost is below the threshold ϵ will be connected by an edge. In this study, to find the significant edges, we fix the edge density of the network at $\rho = \frac{2E}{N(N-1)} = 5\%$ and choose the corresponding threshold $\epsilon(\rho)$ (Malik et al., 2011; Stolbova et al., 2014; Wiedermann et al., 2017).

4.2.3 Network measures

Various network measures are used to quantify the network topology which provide novel insights into the underlying dynamics of the system over different spatial scales (Donges et al., 2009b). We use two network measures to quantify and characterize the spatial pattern of extreme precipitation. One of the simplest local network measure is the *degree* which measures the centrality of a node based on how well-connected it is. The Degree k_i of a node i is defined as

$$k_i = \sum_{j=1}^{N} A_{ij} \tag{4.2}$$

where N is the total number of grid points (nodes). It quantifies the number of direct connections node i has with other nodes in the network (Fig.4.2a)

In climate networks, nodes with higher degree values k_i indicate the spatial distribution of similar variability, related to linked processes (Boers et al., 2014c), such as large-scale atmospheric circulation (Malik et al., 2011; Boers et al., 2013, 2014c). It has been used to identify the highly connected geographical sites (super-nodes) and their association with atmospheric teleconnection pattern (Tsonis et al., 2008; Radebach et al., 2013; Agarwal et al., 2019).

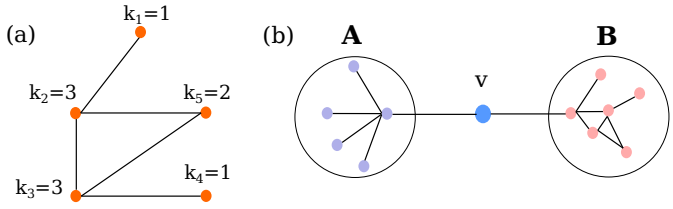


Figure 4.2: Network measures: (a) Degree k_i of the network nodes, based on the number of connections of node i with other nodes. Degree measures how well-connected a node is in the network. (b) Betweenness centrality BC_i of network nodes. Node v has low degree but high betweenness because it acts as a bridge joining two groups of nodes A ans B.

The second network measure we use is the *betweenness centrality*, which provides information about the global topology on the basis of shortest paths between pairs of nodes (Donges et al., 2009b). Betweenness centrality BC_i measures how much a node i falls 'in between' two nodes in the network, i.e., acts as a bridge connecting two other nodes (Newman, 2010; Freeman, 1978). A node may not be well-connected (i.e., has low degree) but can be crucial to connect different parts of the network (Golbeck, 2015) (Fig. 4.2b). Betweenness is quantified by measuring the percentage of the shortest paths that must go through this specific node i and is defined as

$$BC_i = \sum_{j,k \neq i}^{N} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \tag{4.3}$$

where σ_{jk} is the total number of shortest path between node j and k and $\sigma_{jk}(i)$ is the number of shortest paths that go via node i. In case of social network, BC indicates the importance of a node in controlling the flow of information in the network. However, for functional networks, such as climate networks, it represent boundaries between highly connected regions (Molkenthin et al., 2014; Tupikina et al., 2016). BC has been used to uncover energy flow patterns in the atmosphere (Donges et al., 2009a) and has also been successfully applied to study the extreme precipitation patterns of different monsoon systems (Boers et al., 2013; Stolbova et al., 2014).

Correction for spatial embedding: When we choose a particular study area, we impose an artificial boundary in space. These boundaries influence the climate network (Rheinwalt et al., 2012; Boers et al., 2013) by cutting links that actually connect nodes with outer regions, hence affecting the network measures. Here we adopt the boundary correction procedure suggested by (Rheinwalt et al., 2012) as follows: We first generate 500 spatially embedded random networks (Barnett et al., 2007) (SERN) which preserve both the node position and the distribution of the spatial link lengths of the original network. After that, we compute the network measures for all SERN surrogates. The boundary-corrected network measure is obtained by dividing the original measure with the average of the SERN surrogates measure.

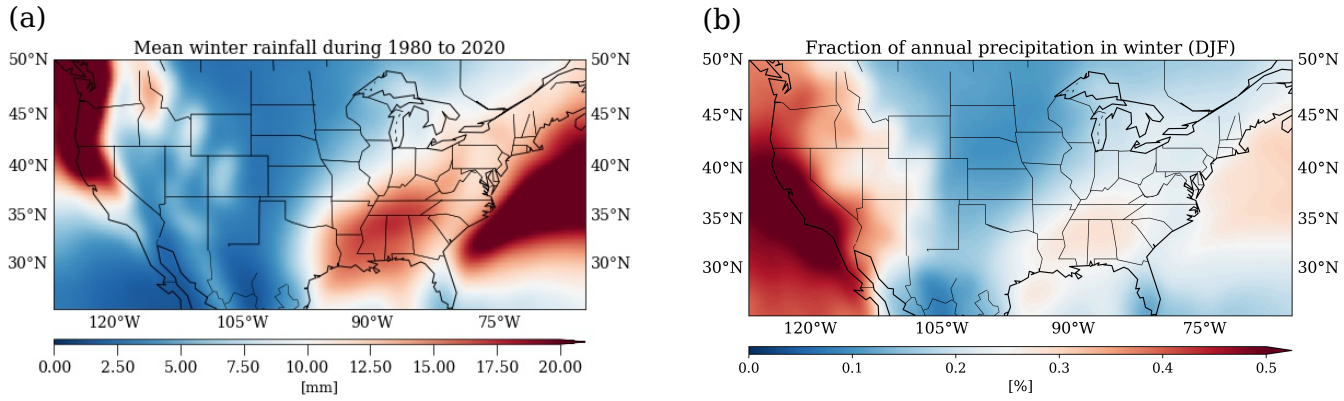


Figure 4.3: (a) Mean daily winter precipitation from 1980 – 2020 (b) Mean winter precipitation anomaly as the fraction of mean annual precipitation falling in winter (same period) for ERA5 reanalysis data. Anomalies are highly positive along the West Coast and slightly positive along the southern flank of the Appalachians. Highly negative anomalies exist in the central north.

4.3 Results and discussion

4.3.1 Calculation and network interpretation

In this section, we analyze the winter extreme precipitation pattern using the above introduced complex network measures. Our climate network, constructed using the ED metric (mentioned in 4.2.2), considers both the sequence and the amplitude of events when quantifying similarity. High degree values (Eq. 4.2) represent high connectivity of extreme precipitation events, i.e., grid cells having similar variation in extreme precipitation occurrence and intensity. We find in the corresponding network a relatively low degree in the northwestern part of the U.S. (Fig. 4.4a), suggesting less similarity of extreme precipitation behaviour with any other regions. A high degree is observed in the eastern Pacific Ocean and southwestern part of the U.S. To understand the connectivity pattern for these regions, we choose really small boxes **A** (low degree) and **B** (high degree), in the northwestern part of the U.S. and in the eastern Pacific Ocean respectively, and determine the number of links connecting these boxes with other

nodes in the network (Figs. 4.5a,b). We find that connections with the region **A** are confined to a very small region centred more towards the coast, indicating a quite narrow corridor of moisture transport as typical for atmospheric rivers (Dettinger, 2013; Xiong and Ren, 2021; Hu et al., 2017; Gonzales et al., 2019). On the other hand, the connectivity of region **B** spans over a larger area in the Pacific Ocean and extending up to some parts of the southwestern coast. Such extended connectivity represents the impact region of a larger atmospheric pattern, such as tropical cyclones which are typical in this region to bring enhanced rainfall (Woodruff et al., 2013). We also observe high degree values in the Great Plains and northeastern parts of the U.S. Here, we choose another small box **C** which lies roughly in the Mississippi river watershed where a relatively high degree is observed (Fig. 4.5c). The connectivity pattern of this region shows a similar behaviour in extreme precipitation along the southwest-northeast direction. It should be noted that high elevation regions such as the Cascades, some parts of the Rockies and the Appalachians show relatively lower degree than low elevation region which was also observed by Agarwal et al. (2022) in case of extreme precipitation networks constructed using edit-distance. Similar observations are made in the results obtained using JRA-55 dataset (see Fig. S2a).

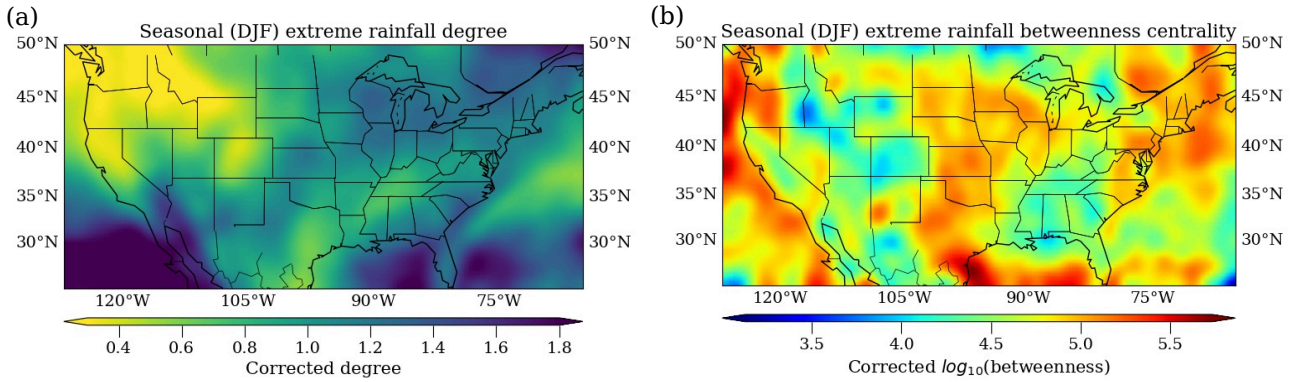


Figure 4.4: (a) Degree, and (b) betweenness centrality for extreme winter (DJF) precipitation from 1980 to 2020.

Next we study the spatial patterns of BC (Fig. 4.4b), representing some striking structures revealing the transition zones between different atmospheric flows (Molkenthin et al., 2014; Tupikina et al., 2016) during winter in the U.S .Along the northwestern coast of the U.S., we find high betweenness but low degree. This implies that although these are relatively small regions of similar precipitation dynamics, they are transition zones of different atmospheric flow directions (Molkenthin et al., 2014), possibly because of spatial confinement and orographic lift due to the presence of topographical features such as mountains. The BC values are seen to continue downwards along the entire western coast, lining the land-sea boundary. The results obtained from ERA5 which is in contrast to those obtained from JRA-55, where they decrease substantially beyond 30°N southwards. We observe high BC values in the central U.S., i.e., from Texas towards the Midwest area, and in the northeastern region, which are also regions of high degree. This implies that while the lower elevation regions, east of the Rocky mountains (Great Plains) and the Appalachians (Coastal Plains), are large regions of spatially coherent extreme precipitation dynamics, big rivers and mountain features cause diversification of atmospheric flow leading to different and strongly fragmented precipitation patterns. These observations are mostly similar with those seen in the network constructed using the JRA-55 dataset (Fig. S2b) except for the small disparity in BC values seen along the southwest coast. This may be due to the relatively high bias in JRA-55 precipitation data in the Pacific ocean close to the tropics (Hassler and Lauer, 2021).

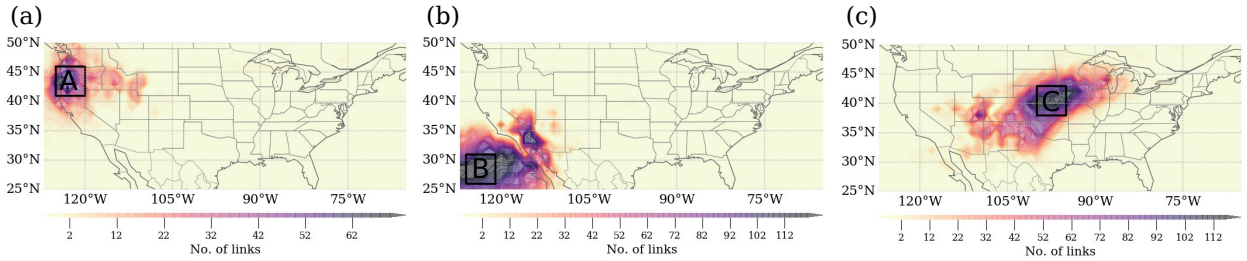


Figure 4.5: Partial degree, i.e., the number links connected to the selected regions in the north-western United States (Box A), eastern Pacific Ocean (Box B), and in the central United States (Box C).

4.3.2 Climatological interpretation

The low connectivity in the northwestern part of the U.S. (Fig.4.4a) is caused by the effects of the Cascade and Rocky Mountains on precipitation. Precipitation gets "trapped" west of these ranges, and, thus, unconnected to the rest of the country, lowering the overall degree. In the high elevations, extreme precipitation requires different conditions than at the coast, so the northwest coast and the mountain ranges are also not connected. However, as the rainstorms can travel more freely through the plains on the eastern side of the mountains, it leads to a higher regional similarity. The presence of the western Cascades results into an orographic lift, effectively transforming the water vapour to extreme precipitation resulting in high BC values little inland up the northwest coast (Fig. 4.4b).

The southwestern part of the U.S. along with adjacent regions of the eastern Pacific Ocean exhibit a high connectivity due to a high fraction of winter precipitation despite a low mean winter precipitation (Fig. 4.3). This effect can be explained by the finding that this part of the eastern Pacific is a separate, relatively small and well-organized precipitation system (Zhang and Wang, 2021) as also seen from Fig. 4.5b. Elevation and slopes are much lower here than further north, so rainstorms can penetrate further into the land and cause near-simultaneous precipitation along the land terrain.

The southwestern coast has high BC values similar to the northwestern coast indicating that they may be related to transition in opposing atmospheric flow direction. The western coast of the U.S. experiences heavy precipitation, and hence extreme streamflows, due to the ARs which contribute 30% to 45% of total winter precipitation (Dettinger, 2013; Xiong and Ren, 2021; Hu et al., 2017). ARs are relatively narrow filament-shaped conduits of moisture in the atmosphere transported from the lower latitudes to the mid and high latitudes (Gimeno et al., 2016; Guan and Waliser, 2015; Ralph et al., 2019). The activity of ARs starts during autumn and tends to shift southward along the Pacific coast later during the winter (Gonzales et al., 2019). However, these ARs may be associated to different regimes of large-scale Rossby wave breaking (RWB) – anticyclonic wave breaking (AWB) in the northwest and cyclonic wave breaking (CWB) in the southwest(Hu et al., 2017)(Fig. 4.6a,b). High BC values penetrate farther inland (Fig. 4.4b), in the northwest U.S., close to the western slope of the Cascades. This may be related to the AWB-ARs which arrive more orthogonally to the western Cascades due to their westerly impinging angle transforming moisture to precipitation due to orographic lift. On the other hand, the CWB-ARs have impinging angles which is more southwesterly, and therefore arrive more orthogonally to the east-west oriented Olympics in the northwest U.S. and the northwest-southeast oriented Sierra Nevada along the southwest coast. Consequently, they causing intense precipitation along the western coast. The transformation of water vapour to extreme precipitation through the orographic lift (Barlow et al., 2019) albeit due to different regimes of RWB explains the high BC along the western coast. The relatively high degree in the southwestern region may be related to the high density of the shorter track ARs close to central and southern California. The seasonal progression of the mean latitude position of the

AR tracks southwards could also possibly explain the high BC values in this region (Gonzales et al., 2019).

The southwest-northeast (SW-NE) inclination in connectivity of the high degree regions in the northeast U.S. and the Great Plains is in agreement with Najibi et al. (2020) who found high similarity in anomalous extreme precipitation in winter in these regions. The eastern side of the Rockies also have high BC values which may be attributed to the pressure gradient seen in the atmosphere (Figs. 4.6a,b,c) (Molkenthin et al., 2014). The area roughly coincides with the loosely-defined region called the Tornado Alley, where tornadoes occur very frequently (Concannon et al., 2000; Bluestein, 2006). We also see a propagation of wind in the southwesterly direction in all atmospheric levels (Figs. 4.6a,b,c). The IVT seasonal composite anomalies (Fig. 4.6d) also show an anomalously high moisture transport in this direction. This flow pattern is modulated by the presence of the Rocky mountains (Lukens et al., 2018) which suppress the storm-track activity by deflecting the westerly flow over land (Chang, 2009). This leads to a SW-NE tilt in the upper tropospheric jet (Fig. 4.6a) subsequently causing a downstream flow and hence high betweenness along those nodes.

High BC values along the northeast coast of U.S. may also be associated with high baroclinic instability formed due to the large land-sea temperature gradient in winter over northeastern U.S.(Brayshaw et al., 2009) which leads to an intensification of extratropical storms on the leeward side of the Appalachian mountains (Colucci, 1976; Lukens et al., 2018). Extreme precipitation in this region is mainly related to an anomalously high upward lift of air along the coast due to high vorticity advection, frequent warm conveyor belts and diabatic heating (Agel et al., 2019). The wind flow (Figs. 4.6a,b,c) and high anomalous IVT (Fig. 4.6d) along the northeast coast, leads to synchronous extreme precipitation in the region and hence high degree.

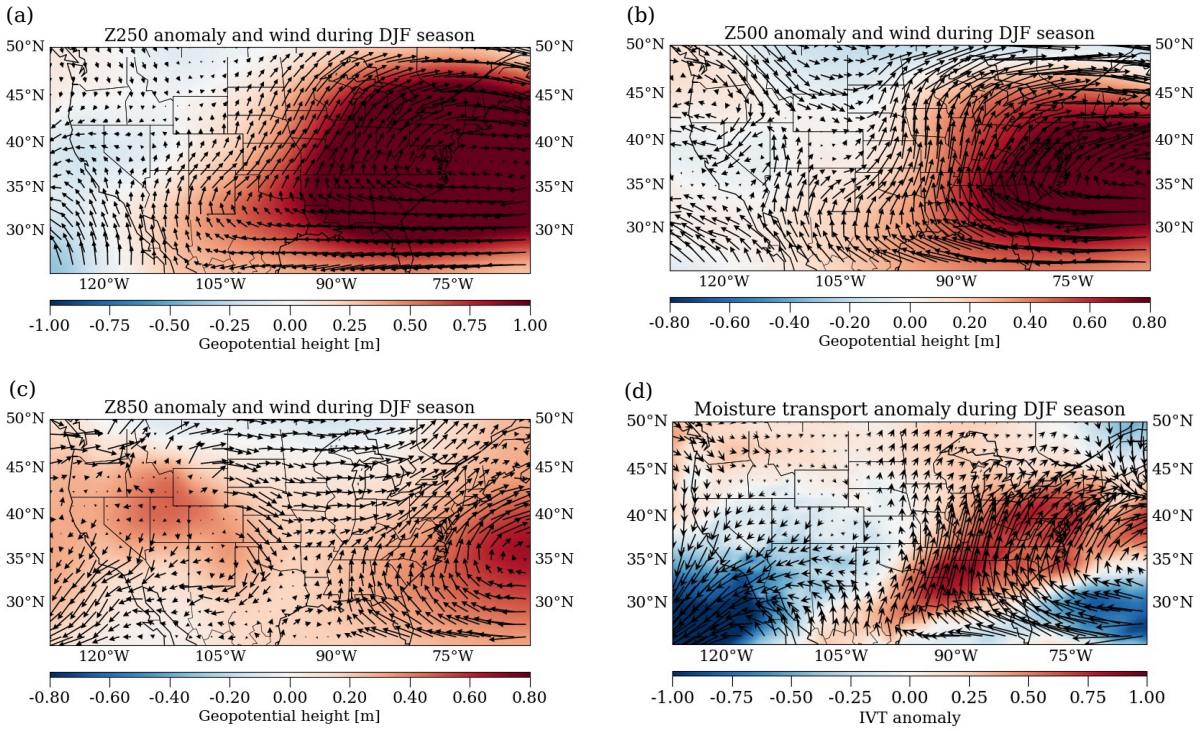


Figure 4.6: (a)-(c) Geopotential height and wind in 250hPa, 500hPa, and 850hPa atmospheric level, (d) vertically integrated water vapour flux anomaly during winter season (DJF).

4.4 Conclusions

The climate network approach has been proven to be a robust and promising framework for studying various climate extremes such as extreme monsoon precipitation (Malik et al., 2011; Boers et al., 2013, 2014c), the influence of El Niño (Boers et al., 2014b), cyclone tracks (Gupta et al., 2021) etc. In this work, we studied the the spatial variability of extreme precipitation during winter for the United States, which has a very complex topography. For this, we employ the edit distance metric to measure pairwise similarity between extreme precipitation time series of different locations. Most of the earlier developed methods (Malik et al., 2011; Stolbova et al., 2014; Boers et al., 2013, 2014c; Wolf et al., 2020) consider only the timing of events when studying the similarities in event-like data. However, the edit distance emerges as a powerful alternative measure because it considers amplitude or strength of extreme events along with their time of occurrences when calculating the similarity.

Extension of the coherent regions depend on the orography, climatological season and the atmospheric circulations. An understanding of the spatial extent of regions of coherent extreme precipitation is necessary for risk assessment of natural hazards. Through a combination of network measures, viz., degree and betweenness centrality, we were able to identify the different regions of the U.S.which exhibit distinctly different extreme precipitation dynamics. While analyzing spatial patterns of degree differentiated between the northwest and the southwest coast on the basis of associated large-scale atmospheric circulation, the high betweenness along the entire western coast brought to light the role of ARs and that of topographic barriers in causing extreme precipitation. The network measures also roughly identify the "Tornado Alley"(Concannon et al., 2000; Bluestein, 2006) region in the Great Plains where tornadoes are more frequent. The high degree pattern captured the southwest-northeast (SW-NE) inclination (Najibi et al., 2020; Lukens et al., 2018) of extreme precipitation due to modulation of storms by the Rocky mountains. Similarly, a modulation of extreme precipitation due to other high ranges, such as the western Cascades and the Appalachians in the east of the country, were also reflected in the network connectivity.

Our complex network based approach provides a comprehensive overview of the distinct regions which experience spatially coherent extreme winter precipitation in the United States albeit due to various climate processes. The similarity measure used in this study, the edit distance, comes out as a very promising alternative to study extreme precipitation patterns in regions exhibiting very intricate climate variability such as the US. Future work may include further refinement of the method by incorporating more sophisticated statistical significance tests. The method can also be applied to study the effects of increasing intensity of extreme precipitation for different large-scale monsoon systems and to possibly identify teleconnections.

Appendix

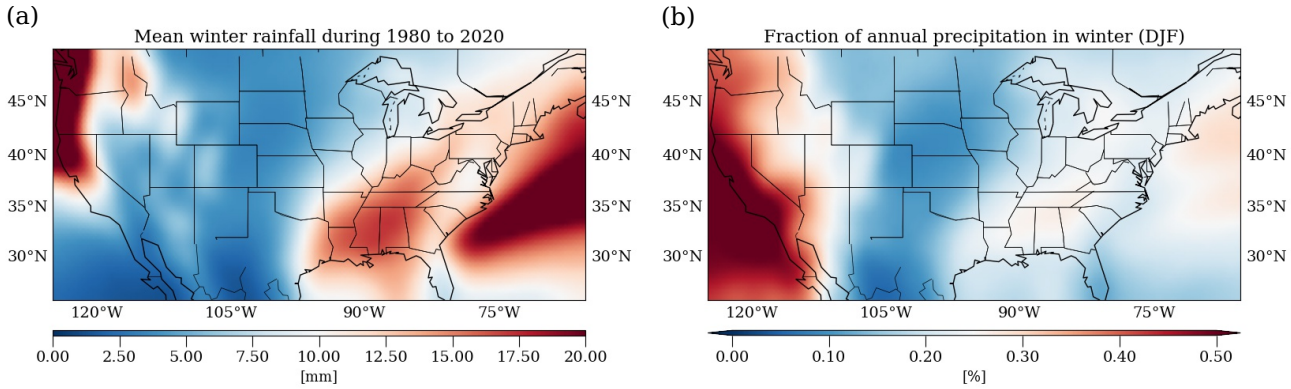


Figure S1: (a) Mean daily winter precipitation from 1980 – 2020 (b) Mean winter precipitation anomaly as the fraction of mean annual precipitation falling in winter (same period) for JRA-55 data.

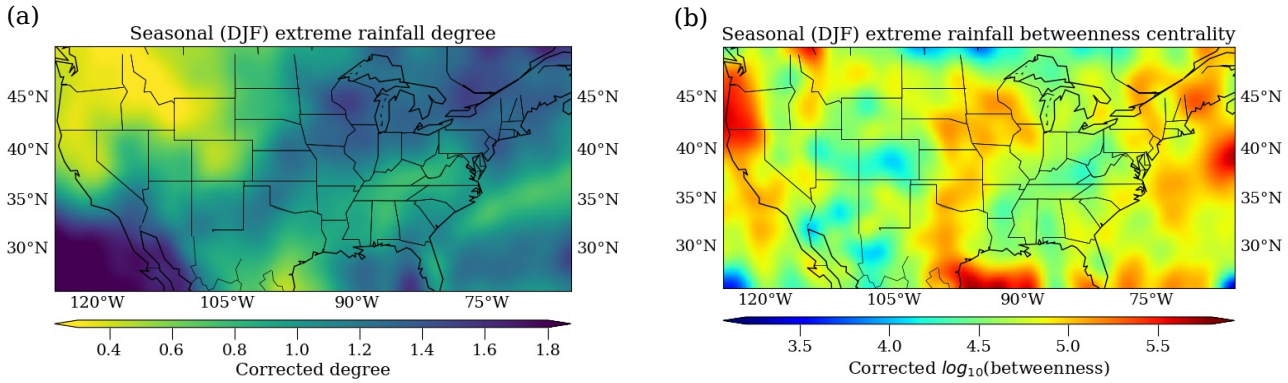


Figure S2: (a) Degree, and (b) betweenness centrality for extreme winter (DJF) precipitation from 1980 to 2020 using JRA 55 reanalysis data.

Chapter 5

Predicting the data structure prior to extreme events from passive observables using the Echo State network

Extreme events are defined as events which largely deviate from the nominal state of the system as observed in a time series. Due to rarity and uncertainty in their occurrence, the prediction of extreme events has been a challenging task. Many a times, in real life, some variables (*passive* variables) encode a significant information about the occurrence of extreme events which are manifested in another variable (*active* variable). For example, often slow varying observables such as temperature, pressure, etc., act as passive variables in case of extreme precipitation events. These passive variables do not show any large excursion from the nominal condition, yet carry the fingerprint of the extreme events. In this study, we propose a reservoir computation based framework that is able to predict the preceding structure or pattern in the time evolution of the active variable that leads to an extreme event using information from the passive variable. An appropriate threshold height of events is a prerequisite for detecting the extreme events and improving the skill of their prediction. We demonstrate that the magnitude of extreme events and the appearance of a coherent pattern before the arrival of the extreme event in a time series, affect the skill of prediction. Quantitatively, we show this using a metric describing the mean phase difference between the input time signals, which decreases when the magnitude of the extreme event is relatively higher, thereby increasing the predictability skill. [1]

[1]Submitted to Frontiers in Applied Mathematics and Statistics (under review) as (P4): Abhirup Banerjee, et al. Predicting the data structure prior to extreme events from passive observables using the Echo State network

5.1 Introduction

In the recent years, extreme events (EEs) have gained attention of researchers and decision makers due to the increase in the occurrence of highly intense climate extremes, such as hurricanes, floods, heatwaves, etc., because of global warming and climate change. They have severe impact on life and infrastructure (Seneviratne et al., 2012; McPhillips et al., 2018; Broska et al., 2020). However, there are several examples of such extraordinary devastating events in various other disciplines aside from climate, like rogue waves in lasers and oceanography, tsunamis and earthquakes in seismology, share market crashes in finance, regime shift in ecosystems, etc.,which are rare but may have long-term correlation in their return periods correlation (Bunde et al., 2003; Jentsch et al., 2006; Dysthe et al., 2008; Altmann et al., 2006; Kharif et al., 2009; Krause et al., 2015b; Marwan and Kurths, 2015; Ray et al., 2020b). The study of extreme events focuses on self-organizing principles (John Rundle, 1996; Sornette, 2002; Jentsch et al., 2006; Chowdhury et al., 2021a; Mishra et al., 2020; Farazmand and Sapsis, 2019) that may enable us to forecast and mitigate the after effect. Various tools have been developed to study the underlying dynamics of such extreme events, e.g., complex network approach has been extensively used to analyze the patterns of climate extremes (Fan et al., 2021b; Boers et al., 2013; Stolbova et al., 2014; Mondal and Mishra, 2021; Agarwal et al., 2022), numerous studies have been conducted to analyze extreme events based on their statistical properties (Ghil et al., 2011; Coles, 2001). Recurrence plot analysis has been used to study the recurring behaviour of flood events (Banerjee et al., 2021). Because of their rare occurrence and complex dynamics, understanding and predicting extreme events is a great challenge in the studies of complex natural systems using the dynamical system approach only (Chowdhury et al., 2021a; Mishra et al., 2020; Karnatak et al., 2014; Ray et al., 2020a; Ansmann et al., 2013). Alternatively, data-based and model-free machine learning techniques have been recently shown to be promising for predicting such events (Amil et al., 2019; Qi and Majda, 2020b; Lellep et al., 2020b; Pyragas and Pyragas, 2020; Chowdhury et al., 2021b; Ray et al., 2021). To put it simply, such a prediction process involves training of the machine using past data records of EEs and then testing the ability of the machine for successfully predicting the prior shape of the observable which leads to the extreme event.

As the term 'extreme event' is used in various disciplines, a precise definition of EEs is not available. Rather it depends on the particular subject where this term is being used. In this work, we select the EEs based on their magnitude. Therefore, it is crucial to set a threshold height so that one can call an event 'extreme 'when it exceeds the threshold. The choice of an appropriate threshold plays a pivotal role in prediction (Hallerberg and Kantz, 2008b,a). In our study, we found that for data-based machine learning, a certain threshold height augments the efficient detection of the arrival of a coherent pattern and thereby leverages the prediction process. In particular, we raise the following question here that for a given multivariate data set in which one of the variables exhibits EEs, whether a seemingly benign variable (with no signature of EE) can be used in a machine for the prediction of the preceding structure or pattern indicative of the forthcoming EE in the malignant (visibly affected) variable.

The aforementioned question is motivated from the fact that occurrence of EEs in one variable are a manifestation of the rich dynamics of a multivariate higher dimensional complex system caused due to the nonlinear interactions among its various constituents (Jentsch et al., 2006; Sornette, 2002). Due to the paucity in observations of naturally occurring EEs, collection or reconstruction of data directly from a dynamic variable that flares up with an extreme value (*active* variable) such as extreme precipitation, over a long time period is seldom possible. However, it is easier to reconstruct data for those observables which are slow varying (temperature, pressure, etc.). Some of these observables may remain silent or *passive* with a slow response to the actual occurrence of events, i.e., do not exhibit extreme values. However,

such *passive* variables carry significant information related to the EEs. We emphasize here that the data collected from the passive variable is used as inputs to a reservoir computing machine, i.e., the echo-state network (ESN), in order to check how efficiently the machine can capture the a priori structure in the *active* observable that precedes the EE. ESN is a simple version of recurrent neural networks (Lukoševičius, 2012) that has been used extensively to predict complex signals ranging from time series generated from chaotic model to stock-price data (Jaeger and Haas, 2004b; Pathak et al., 2018; Zimmermann and Parlitz, 2018; Pathak et al., 2017; Lin et al., 2009; Hinaut and Dominey, 2013; Verstraeten et al., 2005; Lu et al., 2018; Mandal et al., 2022; Lu et al., 2017). Recently, it has been shown that ESN can efficiently capture the onset of generalized synchronization (Weng et al., 2019; Lymburn et al., 2019; Chen et al., 2020; Panday et al., 2021; Fan et al., 2021a), quenching of oscillation (Xiao et al., 2021; Mandal and Shrimali, 2021), detect collective bursting in neuron populations (Saha et al., 2020), and predict epidemic spreading (Ghosh et al., 2021). ESN has been shown to have great potential in handling multiple inputs of temporal data, and ability to trace the relation between them (Lymburn et al., 2019; Saha et al., 2020).

While collecting data, the first important task is to detect the EE by assigning an appropriate threshold height and collect a number of data segments prior to all the available EE in a time series, to address the question of predictability as suggested earlier (Zamora-Munt et al., 2013; Bonatto and Endler, 2017; Hallerberg and Kantz, 2008b). In the present work, we rely on data generated from numerical simulations of a model system for training and testing of the ESN for efficient detection of the structure preceding the extreme events. Firstly, we identify a large number of visible EEs from the active variable using a threshold height and save a data segment of identical length prior to the occurrence of each EE from the active variable along with the corresponding data segment from the passive variable. A multiple number of data segments of identical length corresponding to EEs in the active variable are thus collected from the passive variable and used as inputs to the machine. A part or fraction of the data points from each segment is used for training and the rest of the data points is kept aside for predicting the preceding structure of EE in the active variable during testing.

We repeat the whole process of data collection, training and testing of the machine by varying the choice of the threshold height and then make a quantitative comparison based on predictability skill to select the most suitable threshold height for detection and prediction of EE. It must be noted that by prediction we imply the identification of a common pattern or structure in the test signal (active variable) that always appears quite ahead of time before the arrival of extreme events and hence, effectively works as a precursor to the extreme events. Our machine learning based recipe unfolds two useful information: (i) Data collected from a passive variable before the appearance of EE in an active variable can provide clues to capture the future trend of an active variable and thereby predict the precursory shape of the forthcoming EE, (ii) machine can efficiently suggest a choice of appropriate threshold height that may augment the prediction process. A possible reason for the necessity of a critical threshold for accurate prediction by the machine is explained further in light of a coherent pattern that always appears in the ensemble of multiple segments of data inputs that has been collected prior to the EE.

For demonstration purpose, we use a paradigmatic model neuron that consists of active variables (fast variables) expressing the triggering of extreme events when its passive counterpart (slow variable) shows no signature of extremes.

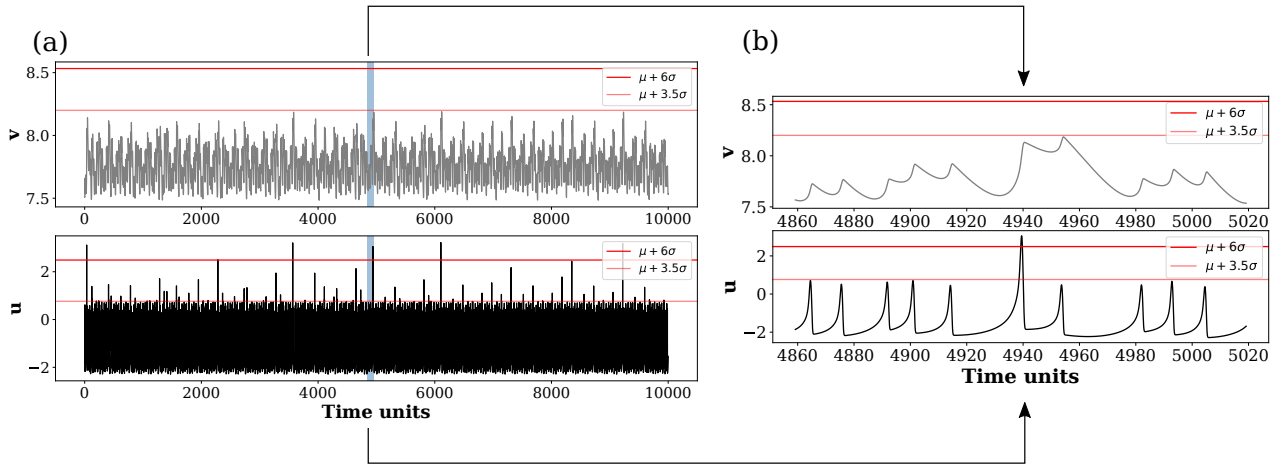


Figure 5.1: Time series of slow variable **v** and fast variable **u** of the coupled Hindmarsh-Rose (HR) system. (a) Horizontal red lines in the time series of u (lower panel) and v (upper panel), indicate two threshold heights $h_{s_1} = \langle\mu\rangle + 3.5\sigma$ (thin line), $h_{s_2} = \langle\mu\rangle + 6\sigma$ (bold line); μ and σ are the mean and standard deviation of the time series, respectively. Threshold height h_{s_2} filters out many large peaks that are otherwise qualified as extremes by the lower threshold h_{s_1}, and thereby allows a selection of rarer extreme events only. One particular extreme peak (shaded region) is marked in (a) as shown in u, and zoomed in the lower panel of (b) for illustration. This extreme peak is larger than both the horizontal lines h_{s_1} and h_{s_2} so as to qualify as a rare extreme event. The corresponding part of the time series of the slow variable v in the upper panel of (a) that never crosses either of the thresholds, h_{s_1} and h_{s_2}, is zoomed in and shown in the upper panel of (b). Although a slight increase in size of the peak is seen (b) compared to its neighboring peaks (upper panel), there is not much significant change in height in comparison to the extreme peak observed in u in the lower panel.

5.2 Methodology

5.2.1 Dataset

For data generation of EEs, we numerically simulate a synaptically (chemically) coupled slow-fast Hindmarsh-Rose (HR) neurons model (Hindmarsh and Rose, 1982),

$$\begin{aligned} \dot{x_i} &= y_i + bx_i^2 - ax_i^3 - z_i + I - \theta_i(x_i - v_s)\Gamma(x_j) \\ \dot{y_i} &= c - \kappa x_i^2 - y_i \\ \dot{z_i} &= r[s(x_i - x_R) - z_i], \end{aligned} \tag{5.1}$$

where x_i and y_i $(i, j = 1, 2;\ i \neq j)$ are the fast variables and oscillate with firing of spiking or bursting potentials, and z_i is the slow variable that controls the fast oscillations; each variable has its specific biological functional meaning. The system parameters a, b, c and s are appropriately chosen where $r < 1$ is the slow parameter. x_R and v_s are constant biases and $\Gamma(x) = \frac{1}{1+\exp^{-\lambda(x-\Theta)}}$ is a sigmoidal function typically used (Mishra et al., 2018) to represent chemical synaptic coupling. The parameters, $a = 1, b = 3, c = 1, \kappa = 5, x_R = -1.6, r = 0.01, s = 5, I = 4, v_s = 2, \lambda = 10, \Theta = -0.25$, are kept fixed for generating data. The coupling constant $\theta_{1,2}$ decides the strength of mutual communication between the neurons via chemical synapses. We collect data on x_i and z_i $(i = 1, 2)$ from numerical simulations and define two new variables, $u = x_1 + x_2$ and $v = z_1 + z_2$. Extreme events are expressed (Mishra et al., 2018) in the fast variable u, which is denoted as our active variable, while the slow variable v is benign and defined as the passive variable. Information from the passive variable v is then used as data input to the machine for predicting the preceding structure of extremes in u.

The local maxima of a time series are identified as events and accordingly all the events are extracted from u for a long run. A standard definition is used for identification of an extreme event (Bonatto et al., 2011; Chowdhury et al., 2021a; Mishra et al., 2020) with a threshold

$h_s = \langle\mu\rangle + d\sigma$, where $\langle\mu\rangle$ is the mean of the long time series, σ is the standard deviation and d is a constant. Any event larger than h_s is considered as an extreme where d is allowed to vary from system to system or for a measured time series under consideration. The question of prediction and enhancing predictability is addressed here by setting different threshold limits h_s by varying d.

For the purpose of numerical experimentation, we first detect a number of extreme peaks n from a long time series of u (total length of the time series : 2×10^7) that crosses a predefined threshold h_s for a particular choice of d. Next, we collect k data points prior to each of the n peaks from u, i.e.,

$$\begin{aligned}
\hat{u}_1 &= (u_1(t), u_1(t-1), u_1(t-2), ..., u_1(t-k)) \\
\hat{u}_2 &= (u_2(t), u_2(t-1), u_2(t-2), ..., u_2(t-k)) \\
&\vdots \\
&\vdots \\
\hat{u}_n &= (u_n(t), u_n(t-1), u_n(t-2), ..., u_n(t-k)),
\end{aligned} \tag{5.2}$$

where $\hat{u}_1, \hat{u}_2, ..., \hat{u}_n$ are the n events selected from the active variable u. We also collect the corresponding data points from the v-time series, i.e.,

$$\begin{aligned}
\hat{v}_1 &= (v_1(t), v_1(t-1), v_1(t-2), ..., v_1(t-k)) \\
\hat{v}_2 &= (v_2(t), v_2(t-1), v_2(t-2), ..., v_2(t-k)) \\
&\vdots \\
&\vdots \\
\hat{v}_n &= (v_n(t), v_n(t-1), v_n(t-2), ..., v_n(t-k)).
\end{aligned} \tag{5.3}$$

In other words, we collect n time segments each containing k data points prior to all the n extreme events, and construct a matrix called event matrix E of size $n \times k$ from the active variable and, similarly, construct a matrix P of the same size $n \times k$ by storing the corresponding data points from the passive variable. A set of m ($m < n$) (grey region A in Fig.5.2a) time segments each with data points p ($p < k$) (Fig.5.2b) as collected from v is then fed into the machine for training to predict the preceding structure of $(n - m)$ segments in u signals (light red region B in Fig.5.2a), which is considered as a precursor to the arrival of extreme events later. How the machine extracts information from the inputs of v and transforms them into u at the output is defined in the input-output functional relation of the machine as a description of the ESN in the next section. Once the training is over, the rest of the $(k-p)$ data points for each of the m time segments are used for testing whether the machine can predict the future structure of $(n - m)$ time segments of u. The whole process is applied multiple times by using four different choices of d (3.5, 4, 5, 6) for detecting extremes from the time series of u. We emphasize once again that an input to the machine for training and testing consists of multiple segments of data points of identical length collected from v corresponding to the successive number of EE detected in u for each d value. The data points collected from u are used at a later stage for comparison with the machine output during the testing process. Certainly this recipe works only when certain amount of data prior to the extreme events is available from both the variables, and the passive variable of the system can be identified. However, the advantage of such a methodology is that it is data-driven and model-free.

5.2.2 Reservoir Computing: Echo-state network model

An echo state network (ESN) is a type of recurrent neural network and is extensively used due to its simple architecture (Lukoševičius, 2012). It has three parts - (1) input layer - in

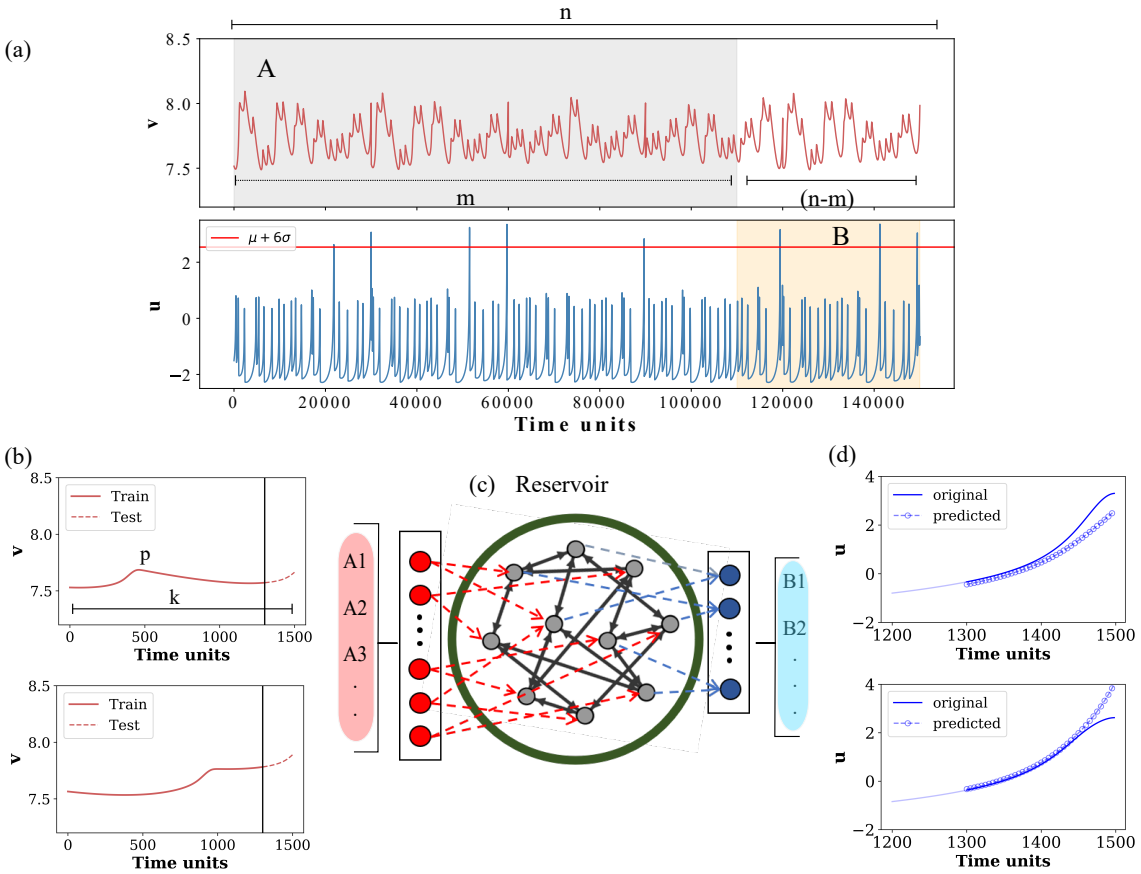


Figure 5.2: Schematic diagram of the ESN and the prediction process. (a) Time series of the passive variable v (upper panel) and active variable u (lower panel) with a number of extreme events, here selected using a threshold height $h_s = \mu + 6\sigma$, are shown. Data points ($k = 1500$) from v, and u prior to n extreme peaks are saved. A few exemplary extreme peaks are shown for demonstration. For our proposed scheme, data points around such $n = 200$ extreme peaks are collected. (b) Two out of the multiple signals corresponding to $m = 180$ extreme events used as inputs for training, are shown as examples. For each input node, $p = 1300$ data points (solid red line) are used for training purpose and the rest of $(k - p) = 200$ data points (dotted red line) are used for testing, which are separated by a vertical line (black line). (c) Echo state network structure: input layer consists of A_m nodes, where $m = 180$ input signals (data segments prior to each of the extreme events) are used for training. The output layer consists of $B_{n-m} = 20$ nodes. (d) Preceding pattern of predicted u signals from 20 nodes each for $(k - p) = 200$ datapoints (blue circles) and the original u signal (blue line) for 200 datapoints are plotted for comparison. Two such output signals are shown as examples.

which the weights are randomly chosen and fixed, (2) reservoir or hidden layer - it is formed by randomly and sparsely connected neurons and (3) output layer - in which the output weights are the only trainable part by input data. A standard leaky network with a tanh activation function is considered here as the ESN. The dynamics of each reservoir node is governed by the following recursive relation:

$$\mathbf{r}(t+1) = (1-\alpha)\mathbf{r}(t) + \alpha \tanh\Big(\mathbf{W}_{res}\mathbf{r}(t) + \mathbf{W}_{in}[1; \mathbf{v}(t)]\Big), \tag{5.4}$$

where $\mathbf{r}(t)$ is a n_{res}-dimensional vector that denotes the state of the reservoir nodes at time instant t, $\mathbf{v}(t)$ is the m-dimensional input vector and 1 is the bias term. The matrices $\mathbf{W}_{res}(n_{res} \times n_{res})$ and $\mathbf{W}_{in}$ $(n_{res} \times (m+1))$ represent the weights of the internal connection of the reservoir nodes and weights of the input, respectively. The parameter α is the leakage constant, which can take any values between 0 to 1. It is to be noted that the tanh function operates element-wise. The choices of α and n_{res} can be varied. Here, we have fixed $\alpha = 0.6$ and $n_{res} = 600$ throughout all simulations. The reservoir weight matrix $\mathbf{W}_{res}$ is constructed by

drawing random numbers uniformly over an interval $[-1, 1]$ and the spectral radius of the matrix $\mathbf{W}_{res}$ is re-scaled to less than unity. The elements of the input weight matrix $\mathbf{W}_{in}$ are also generated randomly from the interval $[-1, 1]$. Next we consider data of n-segments sequentially from the time series of v corresponding to n extreme peaks in u from which a set of first m-segments are fed into the ESN for training. Here p data points (when $t = 0, 1, ..., t_r; p = r + 1$) from the entire length of k data points in each of the m-segments and used for the training purpose (optimizing output weight matrix W_{out}) to capture the trend of the $(n - m)$ segments (each length: $(k - p)$) of u signals. Once the machine is trained, the input of m-segments each with $(k-p)$ data points are fed into the machine to predict the trend of the $(n-m)$-segments of the u signals prior to the arrival of EE in time. At each instant of time t, the $m-$dimensional input vector of data, $\mathbf{v}(t)$:$[v_1(t), v_2(t), ..., v_m(t)]^T$ is fed into m-number of input nodes of the machine when the contribution of the input weight matrix in the dynamics of the reservoir (see Eq. 5.4) is written as,

$$\begin{bmatrix} \mathbf{W}_{in}(1,1) & \cdots & \mathbf{W}_{in}(1,m+1) & \\ \mathbf{W}_{in}(2,1) & \cdots & \mathbf{W}_{in}(2,m+1) & \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{W}_{in}(n_{res},1) & \cdots & \mathbf{W}_{in}(n_{res},m+1) & \end{bmatrix} \times \begin{bmatrix} 1 \\ v_1(t) \\ v_2(t) \\ \vdots \\ v_m(t) \end{bmatrix}.$$

During the training process, at each time instant t, the reservoir state $\mathbf{r}(t)$ and input $\mathbf{v}(t)$ are accumulated in $\mathbf{V}(t) = [1; \mathbf{v}(t); \mathbf{r}(t)]$. The output relation can be written in matrix form,

$$\mathbf{U} = \mathbf{W}_{out}\mathbf{V}, \tag{5.5}$$

where $\mathbf{U}$ is a matrix of dimension $(n - m) \times (k - p)$, and $(k - p)$ is the length of data points in each $(n - m)$ segment. $\mathbf{W}_{out}$ can be determined by the Ridge regression method,

$$\mathbf{W}_{out} = \mathbf{U}\mathbf{V}^T(\mathbf{V}\mathbf{V}^T + \lambda I)^{-1}, \tag{5.6}$$

where λ is the regularization factor that avoids over-fitting. An important point to note is that the output signal from the machine predicts the preceding structure of u prior to the EE in time.

5.3 Results

For illustration of our proposed scheme, the original time series of u and v for a long run of numerical simulations are plotted in Fig. 5.1(a). As the threshold height is increased from $h_{s_1} = \langle\mu\rangle + 3.5\sigma$ and $h_{s_2} = \langle\mu\rangle + 6\sigma$ by varying d from 3.5 to 6, many large peaks are filtered out that declares only a few peaks as rare and extremes. The extreme peaks are selected as those which are higher than a selected threshold height h_s (horizontal line, Fig.5.1(a)) for a particular choice of d, and used as data for training and testing the reservoir shown in Figs. 5.2(b)-(d). It is clear that some of the peaks in u are higher than the designated thresholds h_{s_1} and h_{s_2} whereas the height of all the peaks in v are lower than both thresholds. A zoomed version is shown in Fig. 5.1(b) to demonstrate the time evolution of u and v around a single extreme peak marked by a shaded region in Fig. 5.1(a). Extremes are only expressed in the active variable u with no similar manifestation in the passive variable v, which is considered here as the input candidate to the machine for the prediction of the a priori structure of successive EEs in u.

An exemplary predicted output of u for $(k - p) = 200$ data points (blue circles) vis-à-vis the original u signal of the same length (blue line) is plotted in Fig. 5.3(a)-(d) for four different

d values. A visual impression provides a clear evidence that the error between the predicted signal (blue circles) and the original input signal (blue line) during 1300 to 1500 time units decreases with the increase in the value of d. For a more comprehensive understanding of the

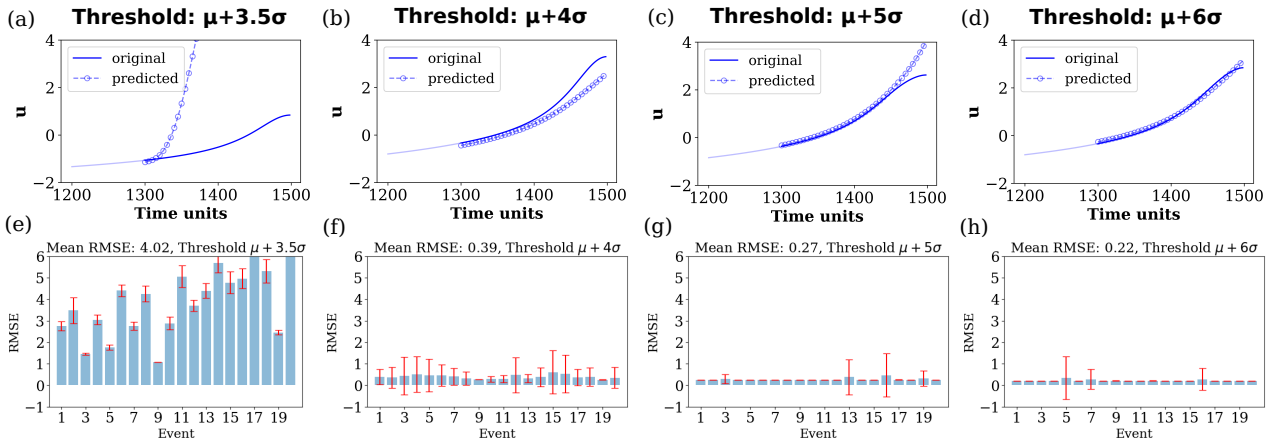


Figure 5.3: Prediction of extreme events by the ESN. Upper panels in (a)-(d) show original active signal u for $(k-p) = 200$ data points (blue line) along with the predicted signal for $(k-p) = 200$ data points (blue circles) for comparison for EEs selected using four different threshold heights computed using: (a) $d = 3.5$, (b) $d = 4$, (c) $d = 5$ and (d) $d = 6$. It shows an increased resemblance between the predicted and original extreme peaks with increasing d. Lower panels in (e)-(h) show RMSE between the original signal u and their predicted signals for $(k-p) = 200$ data during testing, estimated over 20 extreme events, corresponding to (a)-(d) respectively. Results of 400 realizations of data from numerical simulations of the model using 400 different initial conditions for each d value are presented in (e)-(h) and the vertical bars mark their standard deviation.

scenario, the root mean square error (RMSE) estimated for 20 predicted output signals and the original signals of u is plotted which confirms the increasing predictability with higher h_s (Fig. 5.3(e)-(h)). To verify the robustness of the outcome, we repeat the whole process for 400 realizations drawn from 400 different initial conditions. RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{1}{t_f - t_r} \sum_{t=t_r}^{t_f} (u_{original}(t) - u_{machine}(t))^2}. \tag{5.7}$$

where t_r and t_f are training and final time respectively and $t_f - t_r = k - p$.

To understand the reason for the machine's improved performance with a higher h_s, we compare all the 180 input signals of the passive variable (v) as well as the active variable (u) prior to the occurrence of EEs ($p = 1300$ data points) (Fig. 5.4). Upper row plots in Fig. 5.4(a)-(d) represent the input signals v before the EEs for four different threshold values. As we increase the threshold h_s (by increasing d from 3.5, 4, 5, 6), signals observed to get less dispersed and tend to form a coherent bundle. In fact, the increasing coherent pattern among the input signals is more prominent in the corresponding active variable u in the lower row of Fig. 5.4(e)-(h). For the highest threshold value, the time signals are almost coherent similar to what was reported by Zamora-Munt et al. (2013). The increasing coherence in v with higher h_s enhances the machine's predictability skill for higher amplitude events compared to the lower amplitude ones. Thus, the machine establishes a general fact, in quantitative terms, that predictability is enhanced for larger value of threshold height when the input signals are more coherent for a longer duration of time (Zamora-Munt et al., 2013; Bonatto and Endler, 2017).

We repeat our experiments using the same ESN by considering two different length of data inputs ($p = 800, 1300$) prior to each of the extreme events for training, and keeping the same set length of data points $(k - p) = 200$ for testing as done above. The number of inputs

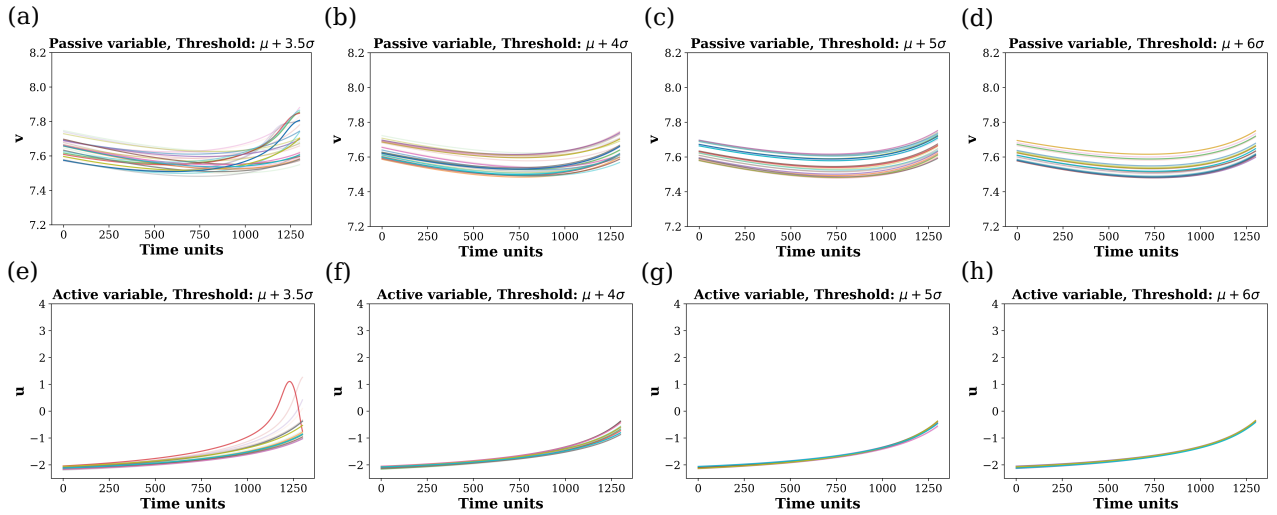


Figure 5.4: Comparative picture of coherence in the input time signals (p) extracted before an extreme events. (a)-(d) Input signal of passive variable v for threshold values ($d = 3.5, 4, 5, 6$). (e)-(h) Output signal of corresponding active variable u for threshold values ($d = 3.5, 4, 5, 6$). Coherence between the input time signals increases with the threshold height determined by higher d values. Different color signifies different trajectories.

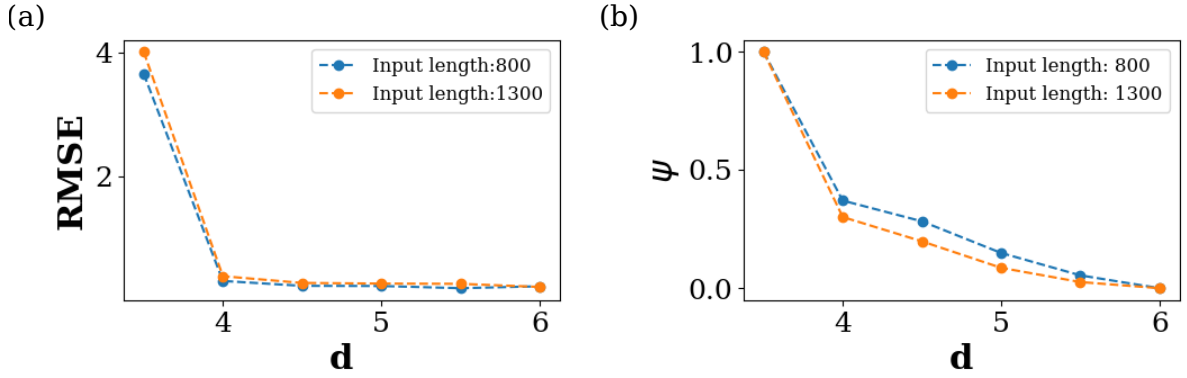


Figure 5.5: Predictability of extreme events. For 20 extreme events, (a) RSME against threshold d for different length of input data, (b) average phase against threshold d for different length of input data. Here, for both cases the average of 400 realizations are presented. Instantaneous phase $\phi_i(t)$ of i^{th} signal is estimated using the Hilbert transform (Rosenblum et al., 1996).

(A_m; $m = 180$) for training and outputs for testing (B_{n-m}; $n - m = 20$) remain unchanged. Thereafter, we calculate the RMSE of the predicted output signals from 20 output nodes for each length of data inputs (p) and repeat the whole process for increasing d values. We plot the RMSE against the d values and for two different time lengths $(800, 1300)$ in Fig. 5.5(a). The RMSE is high for $d = 3.5$, and it gradually decreases and converges to a low value for higher threshold values.

Next we introduce another measure ψ based on the instantaneous phases of the time signal inputs,

$$\psi = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left[\frac{1}{T} \sum_{t=1}^{T} \|\phi_i(t) - \phi_j(t)\| \right] \tag{5.8}$$

where $\phi_i(t)$ is the instantaneous phase of the i-th signal of the passive variable v at time t, n is the total number of segments and T is the segment length. Here, $\phi_i(t)$ of i^{th} signal is calculated using the Hilbert transform (Rosenblum et al., 1996). A High value of ψ indicates less coherent structure and vice-versa. This variable ψ represents the average phase difference (on the number of segment and segment length) between all the 180 input signals of different

length.

We plot values of ψ against d for the two different time lengths $(800, 1300)$ in Fig. 5.5(b). A phase coherence is oberved with increasing d. When the threshold is low (lower value of d) , the time signals of v are dispersed (see Fig. 5.4(a)). As a result, the average phase difference ψ is high. ψ gradually converges for higher values of d with the formation of a coherent bundle of the input signals. This indicates that there is a higher tendency of phase coherence between input signals for higher magnitude EEs which enhances the ability of the machine to predict their precursory structure.

5.4 Conclusion and Discussion

We have proposed an Echo State Network based scheme for the prediction of preceding shape of extreme events from a passive variable which shows no visible manifestation of extreme events, but connected to an active variable that has clear indications of rare and recurrent high amplitude events. Such a situation occurs in the real world where maintaining data records of subsidiary variable is easier, and may be useful for studies related to prediction of extreme events in another observable that is difficult to record. To test our scheme, we generated data using a synaptically (chemical) coupled model of two Hindmarsh-Rose (HR) neurons. Two types of variables are involved in the HR model, two fast variables (defined as active) that exhibit extreme events in their time evolution, and a slow variable (defined here as passive) having a slower time-scale and most importantly, showing no visible signs of extremes. The passive variable was considered as our input candidate for the machine for the purpose of predicting the preceding structure of extreme events in the active variable.

Our strategy was first to identify the extreme events in a long time series of an active variable with a choice of an appropriate threshold height and collect data from the passive variable that corresponds to each extreme in the active variable. We saved the data only prior to the arrival of extreme events barring all extremes, then a part of the collected dataset from the passive variable is used for testing a multi-input machine and another part of the data for testing/predicting the prior structure of the forthcoming extremes. Our results indicated that higher the magnitude of extreme events, the efficiency of the machine to predict its precursory structure is higher. Higher intensity events are defined only by increasing the threshold height. On further investigation, we found that that for higher intensity extreme events the input signals collectively form a coherent pattern, which aided the machine to predict the prior structure with increased efficiency. Thus, coherence of the multi-input time signals is the key to a better prediction of the forthcoming extreme events by the machine. A possible quantitative explanation of the enhanced predictability is provided. For this purpose, a new coherence measure ψ is introduced to represent the average phase differences between the segmented time signals. It was observed that ψ decreases with increasing threshold height, therefore confirming our finding that the enhanced ability of the machine to predict higher amplitude extreme events is related to an increase in the phase coherence of the input signals.

Our machine learning scheme opens up an alternative strategy for predicting extreme events from passive variables in the real world. Furthermore, our findings maintains those reported by Hallerberg and Kantz (2008b,a) that higher the magnitude of extreme events, higher is the predictability skill. Finding suitable passive variables for real world systems is a challenge as they typically belong to very high dimensional system and often can be a combination of multiple variables. As a future scope of this study, we intend to use this procedure to detect and predict the preceding structure of natural extremes, such as floods and earthquakes, which can work as an alert early warning or precursory signal to the forthcoming extreme event.

Chapter 6

Conclusion and outlook

In this thesis, I studied the spatio-temporal patterns of extreme events, and developed a framework to predict the precursory structure of extreme events. While working with extreme events, one comes across some difficulties related to scarcity in the occurrence of these events (big gaps in the time series) which makes it difficult to analyze their spatio-temporal pattern with standard time series analysis tools, in trying to understand the mechanism of the underlying process. To overcome this difficulty, I employed a specific distance metric called the *edit distance* which could handle such data in their unaltered form. To study the temporal and spatial patterns of extreme events, I combined edit distance with well-established methods, namely, recurrence quantification analysis and climate networks, respectively. The application of this methodology was not only successful in studying data from prototypical models but also a number of real-world systems.

In the following, I briefly summarize the interesting results in this thesis obtained from this approach.

- **Temporal patterns of extreme event-like data (Chapter 2)**
  Recurrence plot is an advanced nonlinear data analysis tool, which is used to visually and quantitatively study the temporal dynamics of complex systems. However, standard recurrence plot analysis is not suitable to characterize the dynamics of event-like data. By incorporating edit distance into the framework of recurrence quantification analysis, I accomplished the goal of providing a simple, yet powerful, architecture to study temporal patterns of extreme event-like data. First, I applied this framework to prototypical models, where the governing dynamics are known, which helped me to obtain a vague impression of the recurrence patterns of extreme events observed in data with different underlying processes. Thereafter, I used the method in the real-world example of flood events obtained from the discharge data of the Mississippi river in the USA. A simple visual analysis of the recurrence plot of the flood events revealed the quasi-periodic nature in their occurrence. Using *determinism* (DET) – a popular recurrence quantification measure used to investigate the deterministic property of the data – I found that flood events exhibit significant serial dependency (Hirata and Aihara, 2011). This implies that future events carry some memory from the past. Therefore, edit distance combined with recurrence plot analysis is a promising approach to study the dynamics of extreme event-like data.

- **Spatio-temporal patterns of extreme event-like data (Chapter 3 and Chapter 4)**
  Climate networks are powerful tools used to unravel the spatio-temporal patterns of climate variables such temperature, pressure, rainfall, etc., by encoding information about the statistical interactions between different geographical locations. By characterizing the spatial patterns of the network topology using different network measures such as degree, betweenness centrality, clustering coefficient, etc., one can not only identify the crucial regions in the climate system, but also the climate drivers responsible for spatial coherence

of extreme events, such as extreme rainfall. Although, various nonlinear synchronization measures have been used to study similarity between extreme events, they do not have provision for incorporating the information about the intensity of extreme events. It is here that the novelty of my proposed framework to combine edit-distance with climate network architecture comes into play. First, I used edit distance to study the dynamic relationship between two variables of a known system (section. 3.3.3) with a comparatively large event series. Edit distance showed immense potential in handling large event series and could successfully capture the nonlinear interactions between event-like data. Thereafter, I combined edit distance with the climate network framework (Tsonis and Roebber, 2004; Fan et al., 2021b) to study the collective behavior of extreme rainfall events based on their occurrence and intensity. My application of this methodology to study the extreme precipitation pattern in the Ganga River basin (Chapter 3) revealed the effect of the Himalayan orographic barrier on the precipitation pattern of the Ganges basin. A further application to a comparatively larger study area with complex topologies, like the United States (Chapter 4), unraveled fascinating features about the extreme precipitation pattern during the winter season. For instance, the extreme precipitation pattern on the west coast of the USA is mainly related to the landfall of atmospheric rivers, which may be associated with cyclonic and anticyclonic wave breaking of the Rossby waves (Chapter 4). These two studies showed that edit distance is a highly promising alternative to analyze the functional dependency between extreme event-like data and investigate spatial variability of extreme events under the climate network framework.

Another major challenge arising due to the problem of paucity in observations of extreme events, is predicting them. As we often associate some precursory observations with occurrence of extreme events, prediction of extremes using data-driven methods via identification of precursory signals is an important line of research. It is a crucial premise that the precursory signals may not always be followed by an extreme event, but they should have a high correlation with the event itself. For example, black clouds in the sky indicate high chances of rain but every time, the presence of black clouds in the sky does not necessarily cause rain. Through my previous experience gained from the analysis of extreme events in various prototypical and real-world dynamical systems, I realised that if one variable of a multivariate dynamical system exhibits extreme values (*active*), often other variables of this system with a seemingly slower response (*passive*) can exhibit precursory structures in their data as extreme events are a manifestation of the complex interactions between the different components of the dynamical system. For instance, soil moisture content (*passive* variable) can be early indicators of extreme precipitation (*active* variable) and aid in predictions (Moon and Ha, 2019). Applying this concept of extracting information from passive observables, I built a framework to the predict the data structure a priori to extreme events. The important results from this work is summarized below:

– **Prediction of the preceding structure of extreme events (Chapter 5)**

Using data from prototypical neuron model under a reservoir computing framework, I was able show that the machine can successfully predict precursory structures of extreme events in the active variable when the input data source is derived from the passive observable. Furthermore, we observed the quality of predictions improved when the magnitude of extremes were higher. This was an important result which may be related to higher signal-to-noise ratio for higher magnitude events, i.e., larger events are expected to be better predictable since they and their precursors differ significantly from the typical behavior of the system under study (Hallerberg, 2008). However, this may not sufficient to explain this magnitude dependence. Further investigations revealed that the better

predictability of high intensity extreme events is associated with higher tendency of the input signals to exhibit a phase coherence behaviour.

In my thesis, my work covered a broad range of aspects related to extreme events, from their recurrent nature to synchronization and finally prediction. However, the research field of extreme events is vast, and I understand that my research provides a comparatively narrower perspective.

Further open questions and outlook

In my research on extreme events, I encountered many obstacles, for some of which I found a way through, as discussed above, some I hope to take up as subjects of my future research while some others are still bepuzzling to me. In the context of each part of the thesis, I discuss a few specific open questions below:

While studying the temporal patterns in an extreme event series Chapter 2, finding the optimal window size is still an open question for applying edit distance. A systematic way of choosing the optimum window size that best captures the dynamics of extreme events needs to be devised according to the specific problem. Furthermore, to best exploit the powerful tool of Recurrence plots, more efforts needs to be given to provide physical interpretation of various recurrence quantification measures other than determinism, such as laminarity, recurrence based entropy, etc., (Marwan et al., 2007) in context of extreme events in a specific real-world system.

Next, in the work related to characterization of spatio-temporal dynamics of climate extremes (Chapters 3 and 4) using climate network approach, it must be kept in mind that climate network is a type of functional complex network in which the links represent statistical relationships and not real physical connections. Therefore, it is extremely important to find the significant links which has a high probability of indicating a physical connection via some underlying climate mechanism. Constructing a suitable statistical null model to compute the significance of links is a solution but is quite challenging and computationally expensive. Though, the ultimate test is mostly to find meteorological evidence justifying the presence of the links which is sometimes difficult. The application of edit-distance as a nonlinear synchronization measure in combination with climate network is a very new area of research and requires further in-depth investigations. A possible future scope of this method is to study other kinds of extreme events such as heatwaves, floods, droughts, etc. Moreover, a systematic investigation of the choice of cost parameters for the edit distance measure should be conducted on the basis of characteristics of the event series, such as the length of the time series, the number of events, etc., as real-world extreme event series are characterized by long lengths but sparse events. Furthermore, in my work, I studied the topological properties of climate networks using degree and betweenness centrality. However, there are a plethora of network measures, like eigenvector centrality, clustering coefficient, etc., which can provide further insights into the dynamics of the Earth's climate system about the problem at hand, and therefore should be studied in future.

With reference to my work on prediction of precursory structures of extreme events (Chapter 5), an immediate future scope is to use different machine learning methods to compare the predictability skill. Lastly, an important future problem is to apply the framework to predict extreme events in a real world system, such as floods or earthquakes. However, although there has been works to identify different precursors for these extreme events which can potentially improve forecasting, our lack of complete knowledge of the underlying real-world dynamical system makes this extremely challenging. Given the complexity of the real world systems, it is understandable why methods combining probabilistic models with available precursory information (Shebalin, 2018), prediction based on a multi-parameter approach (Ouzounov et al.,

2018b) in seismology and other fields have failed to provide reliable predictions so far. I am optimistic that increased observations and further advancement of technology, would enhance our ability to observe and better understand these phenomena.

To conclude, the field of extreme events is as diverse as its definition, and therefore it is very difficult to formulate a general framework to cover all aspects of this topic. However, in view of the current scenario, this bepuzzling topic is a doorway to a vast realm of science yet to fully explored and lots of highly significant open questions yet to be solved.

My take home message

Throughout my Ph.D. journey, I was bewildered while trying to connect the various aspects of the diverse topic of 'extreme events' – but "*That's life !*"

At the end, I realised that the more I learnt, the less confident I became because what I know is just a drop of the whole ocean. However, I am deeply motivated to learn more because that one drop of water holds all the secrets of the ocean.

"The woods are lovely, dark and deep
But I have promises to keep
And miles to go before I sleep,
And miles to go before I sleep."
–Robert Frost

Bibliography

Agarwal, A., Caesar, L., Marwan, N., Maheswaran, R., Merz, B., and Kurths, J. (2019). Network-based identification and characterization of teleconnections on different scales. *Scientific Reports*, 9:8808.

Agarwal, A., Guntu, R. K., Banerjee, A., Gadhawe, M. A., and Marwan, N. (2022). A complex network approach to study the extreme precipitation patterns in a river basin. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(1):013113.

Agarwal, A., Marwan, N., Maheswaran, R., Merz, B., and Kurths, J. (2018). Quantifying the roles of single stations within homogeneous regions using complex network analysis. *Journal of Hydrology*, 563:802–810.

Agarwal, A., Marwan, N., Maheswaran, R., Ozturk, U., Kurths, J., and Merz, B. (2020). Optimal design of hydrometric station networks based on complex network analysis. *Hydrology and Earth System Sciences*, 24(5):2235–2251.

Agarwal, A., Marwan, N., Rathinasamy, M., Merz, B., and Kurths, J. (2017). Multi-scale event synchronization analysis for unravelling climate processes: a wavelet-based approach. *Nonlinear Processes in Geophysics*, 24(4):599–611.

Agel, L., Barlow, M., Colby, F., Binder, H., Catto, J., Hoell, A., and Cohen, J. (2019). Dynamical analysis of extreme precipitation in the us northeast based on large-scale meteorological patterns. *Climate Dynamics*, 52:1–22.

Altmann, E. G., Hallerberg, S., and Kantz, H. (2006). Reactions to extreme events: Moving threshold model. *Physica A: Statistical Mechanics and its Applications*, 364:435–444.

Altmann, E. G. and Kantz, H. (2005). Recurrence time analysis, long-term correlations, and extreme events. *Physical Review E*, 71(5):056106.

Amil, P., Soriano, M. C., and Masoller, C. (2019). Machine learning algorithms for predicting the amplitude of chaotic laser pulses. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(11):113111.

Annex, I. (2012). Managing the risks of extreme events and disasters to advance climate change adaptation. *Sciences*, 10:97–104.

Ansmann, G., Karnatak, R., Lehnertz, K., and Feudel, U. (2013). Extreme events in excitable systems and mechanisms of their generation. *Physical Review E*, 88(5):052911.

Arnhold, J., Grassberger, P., Lehnertz, K., and Elger, C. E. (1999). A robust method for detecting interdependences: application to intracranially recorded EEG. *Physica D*, 134(4):419–430.

Babu, G. J. and Feigelson, E. D. (1996). Spatial point processes in astronomy. *Journal of Statistical Planning and Inference*, 50(3):311 – 326. Spatial Statistics, Part I.

Banerjee, A., Goswami, B., Hirata, Y., Eroglu, D., Merz, B., Kurths, J., and Marwan, N. (2021). Recurrence analysis of extreme event-like data. *Nonlinear Processes in Geophysics*, 28(2):213–229.

Barabási, A.-L. (2013). Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375.

Barlow, M., Gutowski, W., Gyakum, J., Katz, R., Lim, Y., Schumacher, R., Wehner, M., Agel, L., Bosilovich, M., Collow, A., Gershunov, A., Grotjahn, R., Leung, R., Milrad, S., and Min, S. (2019). North american extreme precipitation events and related large-scale meteorological patterns: a review of statistical methods, dynamics, modeling, and trends. *Climate Dynamics*, 53:6835 – 6875.

Barnett, L., Di Paolo, E., and Bullock, S. (2007). Spatially embedded random networks. *Phys. Rev. E*, 76:056115.

Barrat, A., Barthelemy, M., and Vespignani, A. (2008). *Dynamical processes on complex networks*. Cambridge university press.

Barros, A. P., Kim, G., Williams, E., and Nesbitt, S. W. (2004). Probing orographic controls in the himalayas during the monsoon using satellite imagery. *Natural Hazards and Earth System Sciences*, 4(1):29–51.

Bastiaans, M. (1985). On the sliding-window representation in digital signal processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(4):868–873.

Bharti, V., Singh, C., Ettema, J., and Turkington, T. (2016). Spatiotemporal characteristics of extreme rainfall events over the northwest himalaya using satellite data. *International Journal of Climatology*, 36(12):3949–3962.

Birkholz, S., Brée, C., Demircan, A., and Steinmeyer, G. (2015). Predictability of rogue events. *Physical review letters*, 114(21):213901.

Bloomfield, P. (2004). *Fourier analysis of time series: an introduction*. John Wiley & Sons.

Blöschl, G., Hall, J., Viglione, A., Perdigão, R. A., Parajka, J., Merz, B., Lun, D., Arheimer, B., Aronica, G. T., Bilibashi, A., et al. (2019). Changing climate both increases and decreases european river floods. *Nature*, 573(7772):108–111.

Bluestein, H. B. (2006). *Tornado alley: monster storms of the Great Plains*. Oxford University Press, USA.

Boers, N., Bookhagen, B., Barbosa, H. M. J., Marwan, N., Kurths, J., and Marengo, J. A. (2014a). Prediction of extreme floods in the eastern Central Andes based on a complex networks approach. *Nature Communications*, 5:5199.

Boers, N., Bookhagen, B., Marwan, N., Kurths, J., and Marengo, J. (2013). Complex networks identify spatial patterns of extreme rainfall events of the South American Monsoon System. *Geophysical Research Letters*, 40(16):4386–4392.

Boers, N., Donner, R., Bookhagen, B., and Kurths, J. (2014b). Complex network analysis helps to identify impacts of the el niño southern oscillation on moisture divergence in south america. *Climate Dynamics*, 45:1–14.

Boers, N., Goswami, B., Rheinwalt, A., Bookhagen, B., Hoskins, B., and Kurths, J. (2019). Complex networks reveal global pattern of extreme-rainfall teleconnections. *Nature*, 566(7744):373–377.

Boers, N., Rheinwalt, A., Bookhagen, B., Barbosa, H. M. J., Marwan, N., Marengo, J., and Kurths, J. (2014c). The south american rainfall dipole: A complex network analysis of extreme events. *Geophysical Research Letters*, 41(20):7397–7405.

Boers, N., Rheinwalt, A., Bookhagen, B., Barbosa, H. M. J., Marwan, N., Marengo, J., and Kurths, J. (2014d). The South American rainfall dipole: A complex network analysis of extreme events. *Geophysical Research Letters*, 41(20):7397–7405.

Bonatto, C. and Endler, A. (2017). Extreme and superextreme events in a loss-modulated co 2 laser: Nonlinear resonance route and precursors. *Physical Review E*, 96(1):012216.

Bonatto, C., Feyereisen, M., Barland, S., Giudici, M., Masoller, C., Leite, J. R. R., and Tredicce, J. R. (2011). Deterministic optical rogue waves. *Physical review letters*, 107(5):053901.

Bradley, E. and Kantz, H. (2015). Nonlinear time-series analysis revisited. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(9):097610.

Brayshaw, D. J., Hoskins, B., and Blackburn, M. (2009). The basic ingredients of the north atlantic storm track. part i: Land?sea contrast and orography. *Journal of the Atmospheric Sciences*, 66(9):2539 – 2558.

Broska, L. H., Poganietz, W.-R., and Vögele, S. (2020). Extreme events definedâĂŤa conceptual discussion applying a complex systems approach. *Futures*, 115:102490.

Brunner, M. I., Gilleland, E., Wood, A., Swain, D. L., and Clark, M. (2020). Spatial dependence of floods shaped by spatiotemporal variations in meteorological and land-surface processes. *Geophysical Research Letters*, 47(13):e2020GL088000. e2020GL088000 10.1029/2020GL088000.

Bunde, A., Eichner, J. F., Havlin, S., and Kantelhardt, J. W. (2003). The effect of long-term correlations on the return periods of rare events. *Physica A: Statistical Mechanics and its Applications*, 330(1-2):1–7.

Buskirk, R. E., Frohlich, C., and Latham, G. V. (1981). Unusual animal behavior before earthquakes: A review of possible sensory mechanisms. *Reviews of geophysics*, 19(2):247–270.

Carrara, A., Guzzetti, F., Cardinali, M., and Reichenbach, P. (1999). Use of gis technology in the prediction and monitoring of landslide hazard. *Natural hazards*, 20(2):117–135.

Chang, E. K. M. (2009). Diabatic and orographic forcing of northern winter stationary waves and storm tracks. *Journal of Climate*, 22(3):670 – 688.

Chelidze, T. and Matcharashvili, T. (2007). Complexity of seismic process; measuring and applications – A review. *Tectonophysics*, 431(1–4):49–60.

Chen, X., Weng, T., Yang, H., Gu, C., Zhang, J., and Small, M. (2020). Mapping topological characteristics of dynamical systems into neural networks: A reservoir computing approach. *Phys. Rev. E*, 102:033314.

Chowdhury, S. N., Ray, A., Dana, S. K., and Ghosh, D. (2021a). Extreme events in dynamical systems and random walkers: A review. *arXiv preprint arXiv:2109.11219*.

Chowdhury, S. N., Ray, A., Mishra, A., and Ghosh, D. (2021b). Extreme events in globally coupled chaotic maps. *Journal of Physics: Complexity*, 2(3):035021.

Ciemer, C., Boers, N., Barbosa, H., Kurths, J., and Rammig, A. (2018). Temporal evolution of the spatial covariability of rainfall in south america. *Climate Dynamics*, 51.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. Springer-Verlag, London.

Colucci, S. J. (1976). winter cyclone frequencies over the eastern united states and adjacent western atlantic, 1964–1973: Student paper-first place winner of the father james b. macelwane annual award in meteorology, announced at the annual meeting of the ams, philadelphia, pa., 21 january 1976. *Bulletin of the American Meteorological Society*, 57(5):548–553.

Comfort, L. K. (2002). Managing intergovernmental responses to terrorism and other extreme events. *Publius: The Journal of Federalism*, 32(4):29–50.

Concannon, P., Brooks, H., and Doswell III, C. (2000). Climatological risk of strong and violent tornadoes in the united states.

Connor, C. B. and Hill, B. E. (1995). Three nonhomogeneous poisson models for the probability of basaltic volcanism: Application to the yucca mountain region, nevada. *Journal of Geophysical Research: Solid Earth*, 100(B6):10107–10125.

Corso, G., Prado, T. D. L., Lima, G. Z. D. S., Kurths, J., and Lopes, S. R. (2018). Quantifying entropy using recurrence matrix microstates. *Chaos*, 28(8):083108.

Cramer, J. (2002). The Origins of Logistic Regression. *Tinbergen Institute Discussion Papers*, 02-119(4).

Crowley, P. M. and Schultz, A. (2010). A new approach to analyzing convergence and synchronicity in growth and business cycles: cross recurrence plots and quantification analysis. *Bank of Finland Research – Discussion Papers*, 16:3–51.

De Haan, L., Ferreira, A., and Ferreira, A. (2006). *Extreme value theory: an introduction*, volume 21. Springer.

Dehghani, A., Sarbishei, O., Glatard, T., and Shihab, E. (2019). A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors. *Sensors (Switzerland)*, 19(22):10–12.

Denny, M. W., Hunt, L. J., Miller, L. P., and Harley, C. D. (2009). On the prediction of extreme ecological events. *Ecological Monographs*, 79(3):397–421.

Desai, B., Maskrey, A., Peduzzi, P., De Bono, A., and Herold, C. (2015). Making development sustainable: the future of disaster risk management, global assessment report on disaster risk reduction.

Dettinger, M. D. (2013). Atmospheric rivers as drought busters on the u.s. west coast. *Journal of Hydrometeorology*, 14(6):1721 – 1732.

Dijkstra, H. A., Hernández-García, E., Masoller, C., and Barreiro, M. (2019). *Networks in Climate*. Cambridge University Press.

Doan, N., Polifke, W., and Magri, L. (2021). Short-and long-term predictions of chaotic flows and extreme events: a physics-constrained reservoir computing approach. *Proceedings of the Royal Society A*, 477(2253):20210135.

Donat, M. G., Lowry, A. L., Alexander, L. V., Ó Gorman, P. A., and Maher, N. (2016). More extreme precipitation in the worldâĂŹs dry and wet regions. *Nature Climate Change*, 6(5):508–513.

Donges, J. F., Schultz, H. C. H., Marwan, N., Zou, Y., and Kurths, J. (2011). Investigating the topology of interacting networks. *European Physical Journal B*, 84:635–651.

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

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.

Donges, J. F., Zou, Y., Marwan, N., and Kurths, J. (2009c). Complex networks in climate dynamics – comparing linear and nonlinear network construction methods. *European Physical Journal – Special Topics*, 174:157–179.

Donner, R. V., Zou, Y., Donges, J. F., Marwan, N., and Kurths, J. (2010). Recurrence networks -a novel paradigm for nonlinear time series analysis. *New Journal of Physics*, 12(3):033025.

Dottori, F., Szewczyk, W., Ciscar, J.-C., Zhao, F., Alfieri, L., Hirabayashi, Y., Bianchi, A., Mongelli, I., Frieler, K., Betts, R. A., et al. (2018). Increased human and economic losses from river flooding with anthropogenic warming. *Nature Climate Change*, 8(9):781–786.

Dysthe, K., Krogstad, H. E., and MÃijller, P. (2008). Oceanic rogue waves. *Annual Review of Fluid Mechanics*, 40(1):287–310.

Easterling, D. R., Kunkel, K. E., Arnold, J. R., Knutson, T., LeGrande, A. N., Leung, L. R., Vose, R. S., Waliser, D. E., and Wehner, M. F. (2017). *Precipitation change in the United States*, pages 207–230. U.S. Global Change Research Program, Washington, D.C.

Ebi, K. L. and Bowen, K. (2016). Extreme events as sources of health vulnerability: Drought as an example. *Weather and climate extremes*, 11:95–102.

Eckmann, J.-P., Kamphorst, S. O., and Ruelle, D. (1987). Recurrence plots of dynamical systems. *Europhysics Letters (EPL)*, 4(9):973–977.

Else, H. (2021). Climate change implicated in germany's deadly floods. *Nature*.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (2013). *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media.

Eroglu, D., McRobie, F. H., Ozken, I., Stemler, T., Wyrwoll, K.-H., Breitenbach, S. F. M., Marwan, N., and Kurths, J. (2016). See-saw relationship of the Holocene East Asian-Australian summer monsoon. *Nature Communications*, 7:12929.

Fan, H., Kong, L.-W., Lai, Y.-C., and Wang, X. (2021a). Anticipating synchronization with machine learning. *Phys. Rev. Res.*, 3(2):023237.

Fan, J., Meng, J., Ludescher, J., Chen, X., Ashkenazy, Y., Kurths, J., Havlin, S., and Schellnhuber, H. J. (2021b). Statistical physics approaches to the complex earth system. *Physics Reports*, 896:1–84. Statistical physics approaches to the complex Earth system.

Farazmand, M. and Sapsis, T. P. (2019). Extreme events: Mechanisms and prediction. *Applied Mechanics Reviews*, 71(5).

Feng, X., Porporato, A., and Rodriguez-Iturbe, I. (2013). Changes in rainfall seasonality in the tropics. *Nature Climate Change*, 3(9):811–815.

Fischer, E., Sippel, S., and Knutti, R. (2021). Increasing probability of record-shattering climate extremes. *Nature Climate Change*, 11(8):689–695.

Flanagan, P. X., Mahmood, R., Umphlett, N. A., Haacker, E., Ray, C., Sorensen, W., Shulski, M., Stiles, C. J., Pearson, D., and Fajman, P. (2020). A hydrometeorological assessment of the historic 2019 flood of nebraska, iowa, and south dakota. *Bulletin of the American Meteorological Society*, 101(6):E817 – E829.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239.

Gadhawe, M. A., Guntu, R. K., and Agarwal, A. (2021). Network-based exploration of basin precipitation based on satellite and observed data. *The European Physical Journal Special Topics*, 230(16):3343–3357.

Gao, Z., Zhang, X., Jin, N., Donner, R. V., Marwan, N., and Kurths, J. (2013). Recurrence networks from multivariate signals for uncovering dynamic transitions of horizontal oil-water stratified flows. *Europhysics Letters*, 103(5):50004.

Gariano, S. L. and Guzzetti, F. (2016). Landslides in a changing climate. *Earth-Science Reviews*, 162:227–252.

Ghil, M., Yiou, P., Hallegatte, S., Malamud, B., Naveau, P., Soloviev, A., Friederichs, P., Keilis-Borok, V., Kondrashov, D., Kossobokov, V., et al. (2011). Extreme events: dynamics, statistics and prediction. *Nonlinear Processes in Geophysics*, 18(3):295–350.

Ghosh, S., Senapati, A., Mishra, A., Chattopadhyay, J., Dana, S. K., Hens, C., and Ghosh, D. (2021). Reservoir computing on epidemic spreading: A case study on covid-19 cases. *Phys. Rev. E*, 104:014308.

Gimeno, L., Dominguez, F., Nieto, R., Trigo, R., Drumond, A., Reason, C. J., Taschetto, A. S., Ramos, A. M., Kumar, R., and Marengo, J. (2016). Major mechanisms of atmospheric moisture transport and their role in extreme precipitation events. *Annual Review of Environment and Resources*, 41(1):117–141.

Golbeck, J. (2015). Chapter 21 - analyzing networks. In Golbeck, J., editor, *Introduction to Social Media Investigation*, pages 221–235. Syngress, Boston.

Gonzales, K. R., Swain, D. L., Nardi, K. M., Barnes, E. A., and Diffenbaugh, N. S. (2019). Recent warming of landfalling atmospheric rivers along the west coast of the united states. *Journal of Geophysical Research: Atmospheres*, 124(13):6810–6826.

Gozolchiani, A., Havlin, S., and Yamasaki, K. (2011). Emergence of el niño as an autonomous component in the climate network. *Physical review letters*, 107(14):148501.

Granese, N. M., Lacapmesure, A., Agüero, M. B., Kovalsky, M. G., Hnilo, A. A., and Tredicce, J. R. (2016). Extreme events and crises observed in an all-solid-state laser with modulation of losses. *Optics letters*, 41(13):3010–3012.

Grassberger, P. (1983). Generalized Dimensions of Strange Attractors. *Physics Letters A*, 97(6):227–230.

Grassberger, P. and Procaccia, I. (1983a). Characterization of strange attractors. *Physical Review Letters*, 50(5):346–349.

Grassberger, P. and Procaccia, I. (1983b). Estimation of the Kolmogorov entropy from a chaotic signal. *Physical Review A*, 9(1–2):2591–2593.

Guan, B. and Waliser, D. E. (2015). Detection of atmospheric rivers: Evaluation and application of an algorithm for global studies. *Journal of Geophysical Research: Atmospheres*, 120(24):12514–12535.

Guisado-Pintado, E. and Jackson, D. W. (2018). Multi-scale variability of storm ophelia 2017: The importance of synchronised environmental variables in coastal impact. *Science of the Total Environment*, 630:287–301.

Guntu, R. K. and Agarwal, A. (2021). Disentangling increasing compound extremes at regional scale during indian summer monsoon. *Scientific reports*, 11(1):1–14.

Guntu, R. K., Maheswaran, R., Agarwal, A., and Singh, V. P. (2020a). Accounting for temporal variability for improved precipitation regionalization based on self-organizing map coupled with information theory. *Journal of Hydrology*, 590:125236.

Guntu, R. K., Rathinasamy, M., Agarwal, A., and Sivakumar, B. (2020b). Spatiotemporal variability of indian rainfall using multiscale entropy. *Journal of Hydrology*, 587:124916.

Guntu, R. K., Yeditha, P. K., Rathinasamy, M., Perc, M., Marwan, N., Kurths, J., and Agarwal, A. (2020c). Wavelet entropy-based evaluation of intrinsic predictability of time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(3):033117.

Gupta, S., Boers, N., Pappenberger, F., and Kurths, J. (2021). Complex network approach for detecting tropical cyclones. *Climate Dynamics*.

Gutschick, V. P. and BassiriRad, H. (2003). Extreme events as shaping physiology, ecology, and evolution of plants: toward a unified definition and evaluation of their consequences. *New Phytologist*, 160(1):21–42.

Hall, J., Arheimer, B., Borga, M., Brázdil, R., Claps, P., Kiss, A., Kjeldsen, T., Kriaučiūnienė, J., Kundzewicz, Z. W., Lang, M., et al. (2014). Understanding flood regime changes in europe: a state-of-the-art assessment. *Hydrology and Earth System Sciences*, 18(7):2735–2772.

Hallerberg, S. (2008). *Predictability of extreme events in time series*. PhD thesis, Universität Wuppertal, Fakultät für Mathematik und Naturwissenschaften .

Hallerberg, S. and Kantz, H. (2008a). How does the quality of a prediction depend on the magnitude of the events under study? *Nonlinear Processes in Geophysics*, 15(2):321–331.

Hallerberg, S. and Kantz, H. (2008b). Influence of the event magnitude on the predictability of an extreme event. *Physical Review E*, 77(1):011108.

Hassler, B. and Lauer, A. (2021). Comparison of reanalysis and observational precipitation datasets including era5 and wfde5. *Atmosphere*, 12(11).

Heckscher, C. M. (2018). A nearctic-neotropical migratory songbird's nesting phenology and clutch size are predictors of accumulated cyclone energy. *Scientific reports*, 8(1):1–6.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

Hinaut, X. and Dominey, P. F. (2013). Real-time parallel processing of grammatical structure in the fronto-striatal system: A recurrent network simulation study using reservoir computing. *PLOS ONE*, 8.

Hindmarsh, J. and Rose, R. (1982). A model of the nerve impulse using two first-order differential equations. *Nature*, 296(5853):162–164.

Hirata, Y. and Aihara, K. (2009). Representing spike trains using constant sampling intervals. *Journal of Neuroscience Methods*, 183(2):277 – 286.

Hirata, Y. and Aihara, K. (2011). Statistical tests for serial dependence and laminarity on recurrence plots. *International Journal of Bifurcation and Chaos*, 21(04):1077–1084.

Hlinka, J., Hartman, D., Vejmelka, M., Novotná, D., and Paluš, M. (2014). Non-linear dependence and teleconnections in climate data: sources, relevance, nonstationarity. *Climate dynamics*, 42(7):1873–1886.

Hoell, A., Hoerling, M., Eischeid, J., and Barsugli, J. (2021). Preconditions for extreme wet winters over the contiguous united states. *Weather and Climate Extremes*, 33:100333.

Houze Jr, R. A. (2012). Orographic effects on precipitating clouds. *Reviews of Geophysics*, 50(1).

Hu, H., Dominguez, F., Wang, Z., Lavers, D. A., Zhang, G., and Ralph, F. M. (2017). Linking atmospheric river hydrological impacts on the u.s. west coast to rossby wave breaking. *Journal of Climate*, 30(9):3381 – 3399.

IPCC, I. (2018). Summary for policymakers "in global warming of 1.5 °c. an ipcc special report on the impacts of global warming of 1.5 °c above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty.

Jaeger, H. (2001). The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, 148(34):13.

Jaeger, H. and Haas, H. (2004a). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *science*.

Jaeger, H. and Haas, H. (2004b). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80.

Janssen, E., Wuebbles, D. J., Kunkel, K. E., Olsen, S. C., and Goodman, A. (2014). Observational- and model-based trends and projections of extreme precipitation over the contiguous united states. *Earth's Future*, 2(2):99–113.

Japan Meteorological Agency, Japan (2013). Jra-55: Japanese 55-year reanalysis, daily 3-hourly and 6-hourly data.

Jentsch, V., Kantz, H., and Albeverio, S. (2006). *Extreme Events: Magic, Mysteries, and Challenges*, pages 1–18. Springer Berlin Heidelberg, Berlin, Heidelberg.

John Rundle, William Klein, D. T. (1996). *RIntroduction*. Addison-Wesley,.

Jongman, B., Hochrainer-Stigler, S., Feyen, L., Aerts, J. C., Mechler, R., Botzen, W. W., Bouwer, L. M., Pflug, G., Rojas, R., and Ward, P. J. (2014). Increasing stress on disaster-risk finance due to large floods. *Nature Climate Change*, 4(4):264–268.

Kalyan, A., Ghose, D. K., Thalagapu, R., Guntu, R. K., Agarwal, A., Kurths, J., and Rathinasamy, M. (2021). Multiscale spatiotemporal analysis of extreme events in the gomati river basin, india. *Atmosphere*, 12(4):480.

Kampen, N. V. (2007). *Chapter II - RANDOM EVENTS*. North-Holland Personal Library. Elsevier, Amsterdam, third edition edition.

Kantz, H. (1994). A robust method to estimate the maximal Lyapunov exponent of a time series. *Physics Letters A*, 185(1):77–87.

Kantz, H. and Schreiber, T. (1997). *Nonlinear Time Series Analysis*. University Press, Cambridge.

Karamanos, K., Dakopoulos, D., Aloupis, K., Peratzakis, A., Athanasopoulou, L., Nikolopoulos, S., Kapiris, P., and Eftaxias, K. (2006). Preseismic electromagnetic signals in terms of complexity. *Physical Review E*, 74(1):016104.

Karnatak, R., Ansmann, G., Feudel, U., and Lehnertz, K. (2014). Route to extreme events in excitable systems. *Physical Review E*, 90(2):022917.

Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in water resources*, 25(8-12):1287–1304.

Kemter, M., Merz, B., Marwan, N., Vorogushyn, S., and Blöschl, G. (2020). Joint trends in flood magnitudes and spatial extents across europe. *Geophysical Research Letters*, 47(7):e2020GL087464.

Kerem, D. H. and Geva, A. B. (2005). Forecasting epilepsy from the heart rate signal. *Medical and biological engineering and computing*, 43(2):230–239.

Kharif, C., Pelinovsky, E., and Slunyaev, A. (2009). *Introduction*, pages 1–10. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kharin, V., Flato, G., Zhang, X., Gillett, N., Zwiers, F., and Anderson, K. (2018). Risks from climate extremes change differently from 1.5 c to 2.0 c depending on rarity. *Earth's Future*, 6(5):704–715.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Konapala, G. and Mishra, A. (2017). Review of complex networks application in hydroclimatic extremes with an implementation to characterize spatio-temporal drought propagation in continental usa. *Journal of Hydrology*, 555:600–620.

Kraemer, K. H., Donner, R. V., Heitzig, J., and Marwan, N. (2018). Recurrence threshold selection for obtaining robust recurrence characteristics in different embedding dimensions. *Chaos*, 28(8):085720.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical review E*, 69(6):066138.

Krause, S. M., Börries, S., and Bornholdt, S. (2015a). Econophysics of adaptive power markets: When a market does not dampen fluctuations but amplifies them. *Physical Review E*, 92(1):012815.

Krause, S. M., Börries, S., and Bornholdt, S. (2015b). Econophysics of adaptive power markets: When a market does not dampen fluctuations but amplifies them. *Phys. Rev. E*, 92:012815.

Kreuz, T., Mormann, F., Andrzejak, R. G., Kraskov, A., Lehnertz, K., and Grassberger, P. (2007). Measuring synchronization in coupled model systems: A comparison of different approaches. *Physica D: Nonlinear Phenomena*, 225(1):29–42.

Kunkel, K. E., Easterling, D. R., Kristovich, D. A. R., Gleason, B., Stoecker, L., and Smith, R. (2012). Meteorological causes of the secular variations in observed extreme precipitation events for the conterminous united states. *Journal of Hydrometeorology*, 13(3):1131 – 1141.

Kursinski, A. L. and Mullen, S. L. (2008). Spatiotemporal variability of hourly precipitation over the eastern contiguous united states from stage iv multisensor analyses. *Journal of Hydrometeorology*, 9(1):3 – 21.

Kurths, J., Agarwal, A., Shukla, R., Marwan, N., Rathinasamy, M., Caesar, L., Krishnan, R., and Merz, B. (2019). Unravelling the spatial diversity of indian precipitation teleconnections via a non-linear multi-scale approach. *Nonlinear Processes in Geophysics*, 26(3):251–266.

Kyrtsou, C. and Vorlow, C. E. (2005). Complex Dynamics in Macroeconomics: A Novel Approach. In Diebolt, C. and Kyrtsou, C., editors, *New Trends in Macroeconomics*, pages 223–238.

Lellep, M., Prexl, J., Linkmann, M., and Eckhardt, B. (2020a). Using machine learning to predict extreme events in the hénon map. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(1):013113.

Lellep, M., Prexl, J., Linkmann, M., and Eckhardt, B. (2020b). Using machine learning to predict extreme events in the hÃĺnon map. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(1):013113.

Lin, X., Yang, Z., and Song, Y. (2009). Short-term stock price prediction based on echo state networks. *Expert Systems with Applications*, 36(3, Part 2):7313–7317.

Loaiciga, H. A. and Mariño, M. A. (1991). Recurrence interval of geophysical events. *Journal of Water Resources Planning and Management*, 117(3):367–382.

Lu, Z., Hunt, B. R., and Ott, E. (2018). Attractor reconstruction by machine learning. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(6):061104.

Lu, Z., Pathak, J., Hunt, B., Girvan, M., Brockett, R., and Ott, E. (2017). Reservoir observers: Model-free inference of unmeasured variables in chaotic systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(4):041102.

Lukens, K. E., Berbery, E. H., and Hodges, K. I. (2018). The imprint of strong-storm tracks on winter weather in north america. *Journal of Climate*, 31(5):2057 – 2074.

Lukoševičius, M. (2012). *A Practical Guide to Applying Echo State Networks*, pages 659–686. Springer Berlin Heidelberg, Berlin, Heidelberg.

Lukoševičius, M. and Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149.

Lymburn, T., Walker, D. M., Small, M., and JÃijngling, T. (2019). The reservoirâĂŹs perspective on generalized synchronization. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(9):093133.

Majumdar, S. N., Pal, A., and Schehr, G. (2020). Extreme value statistics of correlated random variables: a pedagogical review. *Physics Reports*, 840:1–32.

Malik, N., Bookhagen, B., Marwan, N., and Kurths, J. (2011). Analysis of spatial and temporal extreme monsoonal rainfall over south asia using complex networks. *Climate Dynamics*, 39:1–17.

Malik, N., Bookhagen, B., Marwan, N., and Kurths, J. (2012). Analysis of spatial and temporal extreme monsoonal rainfall over south asia using complex networks. *Climate Dynamics*, 39(3–4):971–987.

Malik, N., Marwan, N., and Kurths, J. (2010). Spatial structures and directionalities in Monsoonal precipitation over South Asia. *Nonlinear Processes in Geophysics*, 17(5):371–381.

Mandal, S. and Shrimali, M. D. (2021). Achieving criticality for reservoir computing using environment-induced explosive death. *Chaos*, 31(3):031101.

Mandal, S., Sinha, S., and Shrimali, M. D. (2022). Machine learning potential of a single pendulum. *arXiv preprint arXiv:2201.13390*.

Marwan, N. (2003). *Encounters with neighbours: current developments of concepts based on recurrence plots and their applications*. Norbert Marwan.

Marwan, N. (2008). A historical review of recurrence plots. *The European Physical Journal Special Topics*, 164(1):3–12.

Marwan, N. (2019). *Recurrence plot techniques for the investigation of recurring phenomena in the system earth*. Universität Potsdam.

Marwan, N., Donges, J. F., Zou, Y., Donner, R. V., and Kurths, J. (2009). Complex network approach for recurrence analysis of time series. *Physics Letters A*, 373(46):4246–4254.

Marwan, N. and Kurths, J. (2015). Complex network based techniques to identify extreme events and (sudden) transitions in spatio-temporal systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(9):097609.

Marwan, N., Romano, M. C., Thiel, M., and Kurths, J. (2007). Recurrence Plots for the Analysis of Complex Systems. *Physics Reports*, 438(5–6):237–329.

Marwan, N., Trauth, M. H., Vuille, M., and Kurths, J. (2003). Comparing modern and Pleistocene ENSO-like influences in NW Argentina using nonlinear time series analysis methods. *Climate Dynamics*, 21(3–4):317–326.

Marwan, N., Wessel, N., Meyerfeldt, U., Schirdewan, A., and Kurths, J. (2002). Recurrence Plot Based Measures of Complexity and its Application to Heart Rate Variability Data. *Physical Review E*, 66(2):026702.

McPhillips, L. E., Chang, H., Chester, M. V., Depietri, Y., Friedman, E., Grimm, N. B., Kominoski, J. S., McPhearson, T., Méndez-Lázaro, P., Rosi, E. J., et al. (2018). Defining extreme events: A cross-disciplinary review. *Earth's Future*, 6(3):441–455.

Merz, B., BlÃűschl, G., Vorogushyn, S., Dottori, F., Aerts, J., Bates, P., Bertola, M., Kemter, M., Kreibich, H., Lall, U., and Macdonald, E. (2021). Causes, impacts and patterns of disastrous river floods. *Nature Reviews Earth & Environment*, 2:1–18.

Merz, B., Dung, N. V., Apel, H., Gerlitz, L., Schröter, K., Steirou, E., and Vorogushyn, S. (2018). Spatial coherence of flood-rich and flood-poor periods across germany. *Journal of Hydrology*, 559:813–826.

Miralles, D. G., Teuling, A. J., Van Heerwaarden, C. C., and Vilà-Guerau de Arellano, J. (2014). Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation. *Nature geoscience*, 7(5):345–349.

Mishra, A., Leo Kingston, S., Hens, C., Kapitaniak, T., Feudel, U., and Dana, S. K. (2020). Routes to extreme events in dynamical systems: Dynamical and statistical characteristics. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(6):063114.

Mishra, P. K., Prasad, S., Marwan, N., Anoop, A., Krishnan, R., Gaye, B., Basavaiah, N., Stebich, M., Menzel, P., and Riedel, N. (2018). Contrasting pattern of hydrological changes during the past two millennia from central and northern India: Regional climate differences or anthropogenic impact? *Global and Planetary Change*, 161:97–107.

Mitsui, T. and Boers, N. (2021). Seasonal prediction of indian summer monsoon onset with echo state networks. *Environmental Research Letters*, 16(7):074024.

Mohamad, M. A. and Sapsis, T. P. (2015). Probabilistic description of extreme events in intermittently unstable dynamical systems excited by correlated stochastic processes. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):709–736.

Molkenthin, N., Rehfeld, K., Marwan, N., and Kurths, J. (2014). Networks from Flows – From Dynamics to Topology. *Scientific Reports*, 4(4119):1–5.

Mondal, S. and Mishra, A. K. (2021). Complex networks reveal heatwave patterns and propagations over the usa. *Geophysical Research Letters*, 48(2):e2020GL090411. e2020GL090411 2020GL090411.

Mondal, S., Mishra, A. K., and Leung, L. R. (2020). Spatiotemporal characteristics and propagation of summer extreme precipitation events over united states: A complex network analysis. *Geophysical Research Letters*, 47(15):e2020GL088185. e2020GL088185 2020GL088185.

Moon, S. and Ha, K.-J. (2019). Early indian summer monsoon onset driven by low soil moisture in the iranian desert. *Geophysical Research Letters*, 46(17-18):10568–10577.

Moreno, Y. and Perc, M. (2019). Focus on multilayer networks. *New Journal of Physics*, 22(1):010201.

Murphy, J. M., Sexton, D. M., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., and Stainforth, D. A. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, 430(7001):768–772.

Nadin, M. (2006). Anticipating extreme events. In *Extreme events in nature and society*, pages 21–45. Springer.

Najibi, N., Mazor, A., Devineni, N., Mossel, C., and Booth, J. F. (2020). Understanding the spatial organization of simultaneous heavy precipitation events over the conterminous united states. *Journal of Geophysical Research: Atmospheres*, 125(23):e2020JD033036. e2020JD033036 10.1029/2020JD033036.

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.

Oberst, S. and Tuttle, S. (2018). Nonlinear dynamics of thin-walled elastic structures for applications in space. *Mechanical Systems and Signal Processing*, 110:469–484.

Odenweller, A. and Donner, R. V. (2020). Disentangling synchrony from serial dependency in paired-event time series. *Physical Review E*, 101(5):052213.

Othmer, H. G., Dunbar, S. R., and Alt, W. (1988). Models of dispersal in biological systems. *Journal of mathematical biology*, 26(3):263–298.

Ouzounov, D., Pulinets, S., Hattori, K., and Taylor, P. (2018a). *Pre-earthquake processes: A multidisciplinary approach to earthquake prediction studies*, volume 234. John Wiley & Sons.

Ouzounov, D., Pulinets, S., Liu, J.-Y., Hattori, K., and Han, P. (2018b). Multiparameter assessment of pre-earthquake atmospheric signals. *Pre-earthquake processes: a multidisciplinary approach to earthquake prediction studies*, pages 339–359.

Ozken, I., Eroglu, D., Breitenbach, S. F. M., Marwan, N., Tan, L., Tirnakli, U., and Kurths, J. (2018). Recurrence plot analysis of irregularly sampled data. *Physical Review E*, 98:052215.

Ozken, I., Eroglu, D., Stemler, T., Marwan, N., Bagci, G. B., and Kurths, J. (2015). Transformation-cost time-series method for analyzing irregularly sampled data. *Physical Review E*, 91:062911.

Ozturk, U., Malik, N., Cheung, K., Marwan, N., and Kurths, J. (2019). A network-based comparative study of extreme tropical and frontal storm rainfall over japan. *Climate Dynamics*, 53(1–2):521–532.

Ozturk, U., Marwan, N., Korup, O., Saito, H., Agarwal, A., Grossman, M. J., Zaiki, M., and Kurths, J. (2018). Complex networks for tracking extreme rainfall during typhoons. *Chaos*, 28(7):075301.

Pai, D., Rajeevan, M., Sreejith, O., Mukhopadhyay, B., and Satbha, N. (2014). Development of a new high spatial resolution (0.25× 0.25) long period (1901-2010) daily gridded rainfall data set over india and its comparison with existing data sets over the region. *Mausam*, 65(1):1–18.

Palazzi, E., Von Hardenberg, J., and Provenzale, A. (2013). Precipitation in the hindu-kush karakoram himalaya: observations and future scenarios. *Journal of Geophysical Research: Atmospheres*, 118(1):85–100.

Paluš, M., Komárek, V., Hrnčíř, Z., and Štěrbová, K. (2001). Synchronization as adjustment of information rates: Detection from bivariate time series. *Physical Review E*, 63(4):046211.

Panday, A., Lee, W. S., Dutta, S., and Jalan, S. (2021). Machine learning assisted network classification from symbolic time-series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(3):031106.

Pascale, S., Lucarini, V., Feng, X., Porporato, A., et al. (2015). Analysis of rainfall seasonality from observations and climate models. *Climate Dynamics*, 44(11):3281–3301.

Pathak, J., Hunt, B., Girvan, M., Lu, Z., and Ott, E. (2018). Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Phys. Rev. Lett.*, 120:024102.

Pathak, J., Lu, Z., Hunt, B. R., Girvan, M., and Ott, E. (2017). Using machine learning to replicate chaotic attractors and calculate lyapunov exponents from data. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(12):121102.

Percival, D. B. and Walden, A. T. (2000). *Wavelet methods for time series analysis*, volume 4. Cambridge university press.

Pincus, S. (1995). Approximate entropy (apen) as a complexity measure. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 5(1):110–117.

Poincaré, H. (1890). Sur le problème des trois corps et les équations de la dynamique. *Acta mathematica*, 13(1):A3–A270.

Pyragas, V. and Pyragas, K. (2020). Using reservoir computer to predict and prevent extreme events. *Physics Letters A*, 384(24):126591.

Qi, D. and Majda, A. J. (2020a). Using machine learning to predict extreme events in complex systems. *Proceedings of the National Academy of Sciences*, 117(1):52–59.

Qi, D. and Majda, A. J. (2020b). Using machine learning to predict extreme events in complex systems. *Proceedings of the National Academy of Sciences*, 117(1):52–59.

Quian Quiroga, R., Arnhold, J., and Grassberger, P. (2000). Learning driver-response relationships from synchronization patterns. *Physical Review E*, 61(5):5142–5148.

Quian Quiroga, R., Kraskov, A., Kreuz, T., and Grassberger, P. (2002). Performance of different synchronization measures in real data: A case study on electroencephalographic signals. *Phys. Rev. E*, 65:041903.

Quian Quiroga, R., Kreuz, T., and Grassberger, P. (2002). Event synchronization: A simple and fast method to measure synchronicity and time delay patterns. *Physical Review E*, 66(4):041904.

Radebach, A., Donner, R. V., Runge, J., Donges, J. F., and Kurths, J. (2013). Disentangling different types of el niño episodes by evolving climate network analysis. *Physical Review E*, 88(5):052807.

Raj, S., Shukla, R., Trigo, R. M., Merz, B., Rathinasamy, M., Ramos, A. M., and Agarwal, A. (2021). Ranking and characterization of precipitation extremes for the past 113 years for indian western himalayas. *International Journal of Climatology*, 41(15):6602–6615.

Ralph, F. M., Rutz, J. J., Cordeira, J. M., Dettinger, M., Anderson, M., Reynolds, D., Schick, L. J., and Smallcomb, C. (2019). A scale to characterize the strength and impacts of atmospheric rivers. *Bulletin of the American Meteorological Society*, 100(2):269 – 289.

Rathinasamy, M., Agarwal, A., Sivakumar, B., Marwan, N., and Kurths, J. (2019). Wavelet analysis of precipitation extremes over india and teleconnections to climate indices. *Stochastic Environmental Research and Risk Assessment*, 33(11):2053–2069.

Ray, A., Chakraborty, T., and Ghosh, D. (2021). Optimized ensemble deep learning framework for scalable forecasting of dynamics containing extreme events. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(11):111105.

Ray, A., Mishra, A., Ghosh, D., Kapitaniak, T., Dana, S. K., and Hens, C. (2020a). Extreme events in a network of heterogeneous josephson junctions. *Physical Review E*, 101(3):032209.

Ray, A., Rakshit, S., Basak, G. K., Dana, S. K., and Ghosh, D. (2020b). Understanding the origin of extreme events in el niño southern oscillation. *Physical Review E*, 101(6):062210.

Rheinwalt, A., Marwan, N., Kurths, J., Werner, P., and Gerstengarbe, F.-W. (2012). Boundary effects in network measures of spatially embedded networks. In *2012 SC Companion: High Performance Computing, Networking Storage and Analysis*, pages 500–505. IEEE.

Rings, T., Mazarei, M., Akhshi, A., Geier, C., Tabar, M., and Lehnertz, K. (2019a). Traceability and dynamical resistance of precursor of extreme events. *Scientific reports*, 9(1):1–10.

Rings, T., von Wrede, R., and Lehnertz, K. (2019b). Precursors of seizures due to specific spatial-temporal modifications of evolving large-scale epileptic brain networks. *Scientific reports*, 9(1):1–12.

Rodó, X. and Rodríguez-Arias, M.-A. (2006). A new method to detect transitory signatures and local time/space variability structures in the climate system: the scale-dependent correlation analysis. *Climate Dynamics*, 27(5):441–458.

Romano, M. C., Thiel, M., Kurths, J., and Grebogi, C. (2007). Estimation of the direction of the coupling by conditional probabilities of recurrence. *Phys. Rev. E*, 76:036211.

Rosenblum, G., Pikovsky, A. S., and Kurths, J. (1997). From Phase to Lag Synchronization in Coupled Chaotic Oscillators. *Physical Review Letters*, 78(22):4193–4196.

Rosenblum, M. G., Pikovsky, A. S., and Kurths, J. (1996). Phase synchronization of chaotic oscillators. *Physical review letters*, 76(11):1804.

Rosenstein, M. T., Collins, J. J., and De Luca, C. J. (1993). A practical method for calculating largest Lyapunov exponents from small data sets. *Physica D: Nonlinear Phenomena*, 65(1-2):117–134.

Ross, S. M. (1997). *Introduction to Probability Models*. Academic Press, San Diego, CA, USA, sixth edition.

Runge, J., Heitzig, J., Marwan, N., and Kurths, J. (2012). Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy. *Physical Review E*, 86:061121.

Saha, S., Mishra, A., Ghosh, S., Dana, S. K., and Hens, C. (2020). Predicting bursting in a complete graph of mixed population through reservoir computing. *Phys. Rev. Research*, 2:033338.

Sakoe, H. and Chiba, S. (1978). Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-26(1):43–49.

Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., Van Nes, E. H., Rietkerk, M., and Sugihara, G. (2009). Early-warning signals for critical transitions. *Nature*, 461(7260):53–59.

Scheinkman, J. A. and LeBaron, B. (1989). Nonlinear dynamics and GNP data. In *Economic complexity: chaos, sunspots, bubbles, and nonlinearity*, pages 213–227. Cambridge University Press, Cambridge.

Schiff, S. J., So, P., Chang, T., Burke, R. E., and Sauer, T. (1996). Detecting dynamical interdependence and generalized synchrony through mutual prediction in a neural ensemble. *Phys. Rev. E*, 54:6708–6724.

Schinkel, S., Marwan, N., Dimigen, O., and Kurths, J. (2009). Confidence bounds of recurrence-based complexity measures. *Physics Letters A*, 373(26):2245–2250.

Schröter, K., Kunz, M., Elmer, F., Mühr, B., and Merz, B. (2015). What made the june 2013 flood in germany an exceptional event? a hydro-meteorological evaluation. *Hydrology and Earth System Sciences*, 19(1):309–327.

Seneviratne, S., Nicholls, N., Easterling, D., Goodess, C., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., et al. (2012). Changes in climate extremes and their impacts on the natural physical environment.

Shebalin, P. (2018). Combining probabilistic seismicity models with precursory information. *Pre-Earthquake Processes: A Multidisciplinary Approach to Earthquake Prediction Studies*, 234:173.

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*, pages 517–524.

Sivakumar, B. and Woldemeskel, F. M. (2014). Complex networks for streamflow dynamics. *Hydrology and Earth System Sciences*, 18(11):4565–4578.

Smith, R. L. (2003). Statistics of extremes, with applications in environment, insurance and finance. *Extreme values in finance, telecommunications and the environment*, 1:78.

Sornette, D. (2002). Predictability of catastrophic events: Material rupture, earthquakes, turbulence, financial crashes, and human birth. *Proceedings of the National Academy of Sciences*, 99(Supplement 1):2522âĂŞ2529.

Stangalini, M., Ermolli, I., Consolini, G., and Giorgi, F. (2017). Recurrence quantification analysis of two solar cycle indices. *Journal of Space Weather and Space Climate*, 7(a5):1–13.

Stolbova, V., Martin, P., Bookhagen, B., Marwan, N., and Kurths, J. (2014). Topology and seasonal evolution of the network of extreme precipitation over the indian subcontinent and sri lanka. *Nonlinear Processes in Geophysics*, 21(4):901–917.

Strozzi, F., Zaldívar, J.-M., and Zbilut, J. P. (2002). Application of nonlinear time series analysis techniques to high-frequency currency exchange data. *Physica A*, 312(3–4):520–538.

Sura, P. (2011). A general perspective of extreme events in weather and climate. *Atmospheric Research*, 101(1-2):1–21.

Suzuki, S., Hirata, Y., and Aihara, K. (2010). Definition of distance for marked point process data and its application to recurrence plot-based analysis of exchange tick data of foreign currencies. *International Journal of Bifurcation and Chaos*, 20(11):3699–3708.

Swierczynski, T., Lauterbach, S., Dulski, P., Delgado, J., Merz, B., and Brauer, A. (2013). Mid- to late holocene flood frequency changes in the northeastern alps as recorded in varved sediments of lake mondsee (upper austria). *Quaternary Science Reviews*, 80:78 – 90.

Tabari, H. (2020). Climate change impact on flood and extreme precipitation increases with water availability. *Scientific Reports*, 10.

Tabari, H. (2021). Extreme value analysis dilemma for climate change impact assessment on global flood and extreme precipitation. *Journal of Hydrology*, 593:125932.

Takens, F. (1981). Detecting Strange Attractors in Turbulence. In Rand, D. and Young, L.-S., editors, *Dynamical Systems and Turbulence*, volume 898 of *Lecture Notes in Mathematics*, pages 366–381. Springer, Berlin.

Tang, Y., Kurths, J., Lin, W., Ott, E., and Kocarev, L. (2020). Introduction to focus issue: When machine learning meets complex systems: Networks, chaos, and nonlinear dynamics. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(6):063151.

Thayyen, R. J., Dimri, A., Kumar, P., and Agnihotri, G. (2013). Study of cloudburst and flash floods around leh, india, during august 4–6, 2010. *Natural hazards*, 65(3):2175–2204.

Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., and Farmer, B. (1992). Testing for nonlinearity in time series: the method of surrogate data. *Physica D*, 58:77–94.

Thompson, H. (1955). Spatial point processes, with applications to ecology. *Biometrika*, 42(1/2):102–115.

Touma, D., Michalak, A. M., Swain, D. L., and Diffenbaugh, N. S. (2018). Characterizing the spatial scales of extreme daily precipitation in the united states. *Journal of Climate*, 31(19):8023 – 8037.

Trauth, M. H., Asrat, A., Duesing, W., Foerster, V., Kraemer, K. H., Marwan, N., Maslin, M. A., and Schaebitz, F. (2019). Classifying past climate change in the Chew Bahir basin, southern Ethiopia, using recurrence quantification analysis. *Climate Dynamics*, 53(5):2557–2572.

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

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.

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.

Tupikina, L., Molkenthin, N., López, C., Hernández-García, E., Marwan, N., and Kurths, J. (2016). Correlation Networks from Flows. The Case of Forced and Time-Dependent Advection-Diffusion Dynamics. *PLoS ONE*, 11(4):e0153703.

Van Oldenborgh, G. J., Van Der Wiel, K., Sebastian, A., Singh, R., Arrighi, J., Otto, F., Haustein, K., Li, S., Vecchi, G., and Cullen, H. (2017). Attribution of extreme rainfall from hurricane harvey, august 2017. *Environmental Research Letters*, 12(12):124009.

Vejmelka, M. and Paluš, M. (2008). Inferring the directionality of coupling with conditional mutual information. *Physical Review E*, 77(2):026214.

Verstraeten, D., Schrauwen, B., and Stroobandt, D. (2006). Reservoir-based techniques for speech recognition. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 1050–1053. IEEE.

Verstraeten, D., Schrauwen, B., Stroobandt, D., and Van Campenhout, J. (2005). Isolated word recognition with the liquid state machine: a case study. *Information Processing Letters*, 95(6):521–528. Applications of Spiking Neural Networks.

Victor, J. D. and Purpura, K. P. (1997). Metric-space analysis of spike trains: theory, algorithms and application. *Network: Computation in Neural Systems*, 8(2):127–164.

Vu, T. M. and Mishra, A. K. (2019). Nonstationary frequency analysis of the recent extreme precipitation events in the united states. *Journal of Hydrology*, 575:999–1010.

Walleshauser, B. and Bollt, E. (2022). Predicting sea surface temperatures with coupled reservoir computers. *Nonlinear Processes in Geophysics Discussions*, pages 1–19.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of âĂŸsmall-worldâĂŹnetworks. *nature*, 393(6684):440–442.

Weather, B.-D. and for Environmental Information (NCEI), C. D. . N. C. Summary stats | billion-dollar weather and climate disasters | national centers for environmental information (ncei). (Accessed on 10/11/2021).

Webb, G. R. (2002). Sociology, disasters, and terrorism: Understanding threats of the new millennium. *Sociological focus*, 35(1):87–95.

Webber, Jr., C. L. and Zbilut, J. P. (1994). Dynamical assessment of physiological systems and states using recurrence plot strategies. *Journal of Applied Physiology*, 76(2):965–973.

Wendi, D., Merz, B., and Marwan, N. (2019). Assessing Hydrograph Similarity and Rare Runoff Dynamics by Cross Recurrence Plots. *Water Resources Research*, 55(6):4704–4726.

Weng, T., Yang, H., Gu, C., Zhang, J., and Small, M. (2019). Synchronization of chaotic systems and their machine-learning models. *Phys. Rev. E*, 99:042203.

Wernli, H., Dirren, S., Liniger, M. A., and Zillig, M. (2002). Dynamical aspects of the life cycle of the winter storm 'lothar'(24–26 december 1999). *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 128(580):405–429.

Wiedermann, M., Donges, J. F., Kurths, J., and Donner, R. V. (2017). Mapping and discrimination of networks in the complexity-entropy plane. *Phys. Rev. E*, 96:042304.

Wolf, A., Swift, J. B., Swinney, H. L., and Vastano, J. A. (1985). Determining Lyapunov Exponents from a Time Series. *Physica D*, 16(3):285–317.

Wolf, F., Bauer, J., Boers, N., and Donner, R. V. (2020). Event synchrony measures for functional climate network analysis: A case study on south american rainfall dynamics. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(3):033102.

Wolf, F., Ozturk, U., Cheung, K., and Donner, R. (2021). Spatiotemporal patterns of synchronous heavy rainfall events in east asia during the baiu season. *Earth System Dynamics*, 12:295–312.

Woodruff, J., Irish, J., and Camargo, S. (2013). Coastal flooding by tropical cyclones and sea-level rise. *Nature*, 504:44–52.

Wulf, H., Bookhagen, B., and Scherler, D. (2010). Seasonal precipitation gradients and their impact on fluvial sediment flux in the northwest himalaya. *Geomorphology*, 118(1-2):13–21.

Xiao, R., Kong, L.-W., Sun, Z.-K., and Lai, Y.-C. (2021). Predicting amplitude death with machine learning. *Phys. Rev. E*, 104(1):014205.

Xiong, Y. and Ren, X. (2021). Influences of atmospheric rivers on north pacific winter precipitation: Climatology and dependence on enso condition. *Journal of Climate*, 34(1):277 – 292.

Yadav, B. C., Thayyen, R. J., and Jain, K. (2020). Topoclimatic zones and characteristics of the upper ganga basin, uttarakhand, india. *International Journal of Climatology*, 40(14):6002–6019.

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.

Zamora-Munt, J., Garbin, B., Barland, S., Giudici, M., Leite, J. R. R., Masoller, C., and Tredicce, J. R. (2013). Rogue waves in optically injected lasers: Origin, predictability, and suppression. *Phys. Rev. A*, 87:035802.

Zbilut, J. P., Thomasson, N., and Webber, Jr., C. L. (2002). Recurrence quantification analysis as a tool for nonlinear exploration of nonstationary cardiac signals. *Medical Engineering & Physics*, 24(1):53–60.

Zhang, Y. and Wang, K. (2021). Global precipitation system size. *Environmental Research Letters*, 16(5):054005.

Zhou, C., Zemanová, L., Zamora, G., Hilgetag, C. C., and Kurths, J. (2006). Hierarchical organization unveiled by functional connectivity in complex brain networks. *Physical review letters*, 97(23):238103.

Zimmermann, R. S. and Parlitz, U. (2018). Observing spatio-temporal dynamics of excitable media using reservoir computing. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(4):043118.