2007

Nonlinear dependence and extremes in hydrology and climate

Shiraj Khan
University of South Florida

Nonlinear Dependence and Extremes in Hydrology and Climate

by

Shiraj Khan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Civil and Environmental Engineering
College of Engineering
University of South Florida

Co-Major Professor: Auroop R. Ganguly, Ph.D.
Co-Major Professor: Sunil Saigal, Ph.D.
David J. Erickson III, Ph.D.
Thomas Wilbanks, Ph.D.
Manish Agrawal, Ph.D.
Bellie Sivakumar, Ph.D.

Date of Approval:
June 22, 2007

Keywords: Mutual information, South America, Precipitation, Time series, Extreme value distribution, CCSM3 climate model, Chaos

Dedication

To My Family and all Whom I Love

Acknowledgments

This dissertation study was partially funded by the SEED money funds of the Laboratory Directed Research and Development Program of the Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U.S. Department of Energy (DOE) under Contract No. DE-AC05-00OR22725. The SEED money funded research has resulted in published or accepted peer-reviewed papers. Thus, this dissertation study is not subject to export control or related U.S. DOE regulations.

I wish to express my special gratitude to both of my mentors Dr. Auroop R. Ganguly and Prof. Sunil Saigal for generously supporting my dissertation with their time, commitment, scientific expertise, and helpful comments and suggestions. I would like to thank them from the bottom of my heart for encouraging me and stimulating my analytical and scientific thinking throughout my doctoral work which helped me immensely to develop independent thinking, research and writing skills. I would also like to thank Prof. Sunil Saigal for helping me realize my dreams of pursuing higher studies in the United States by offering me full scholarship in the Department of Civil and Environmental Engineering at Carnegie Mellon University (CMU) and for continually supporting me financially during my doctoral work.

I am grateful for having an exceptional doctoral committee and wish to thank Dr. David J. Erickson III, Dr. Thomas Wilbanks, Prof. Manish Agrawal, and Prof. Bellie Sivakumar for their continual support and encouragement. I wish to acknowledge and thank Drs. David J. Erickson III, Sharba Bandyopadhyay, Gabriel Kuhn, George Ostrouchov, Marcia Branstetter, and Prof. Amar Gupta for collaborating and imparting knowledge at various stages of my dissertation. I would like to thank all the people who provided feedback and support at various phases of my dissertation: Profs. Guiling Wang, Shafiqul Islam, Norberto O. Garcìa, Rafael Bras, Tailen Hsing, and Drs. Alexander Kraskov, Earle Williams, Carlos Nobre, Rick Katz, Brant Liebmann, Dave Allured, United States Geological Survey (USGS) and Bureau of Reclamation (BoR). I would also like to take this opportunity to thank all my teachers at IIT Roorkee, CMU, and USF. My special thanks also go out to all my friends and apartment-mates in Tampa, Knoxville, Pittsburgh and elsewhere for supporting, encouraging and bearing with me throughout my doctoral studies.

Finally, I am very grateful to my mother (Rajia Khan) and father (Basi Ullah Khan) for their unwavering faith in my abilities, showing me the true worth of hard work, and supporting me throughout my life. I would like to thank Firoj (brother), Farha (sister), and Nazim (brother-in-law) for their love, support, and encouragement.

Table of Contents

List of Tables

List of Figures

Nonlinear Dependence and Extremes in Hydrology and Climate

Shiraj Khan

ABSTRACT

The presence of nonlinear dependence and chaos has strong implications for predictive modeling and the analysis of dominant processes in hydrology and climate. Analysis of extremes may aid in developing predictive models in hydro-climatology by giving enhanced understanding of processes driving the extremes and perhaps delineate possible anthropogenic or natural causes. This dissertation develops and utilizes different set of tools for predictive modeling, specifically nonlinear dependence, extreme, and chaos, and tests the viability of these tools on the real data. Commonly used dependence measures, such as linear correlation, cross-correlogram or Kendall's τ, cannot capture the complete dependence structure in data unless the structure is restricted to linear, periodic or monotonic. Mutual information (MI) has been frequently utilized for capturing the complete dependence structure including nonlinear dependence. Since the geophysical data are generally finite and noisy, this dissertation attempts to address a key gap in the literature, specifically, the evaluation of recently proposed MI-estimation methods to choose the best method for capturing nonlinear dependence, particularly in terms of their robustness for short and noisy data. The performance of kernel density estimators (KDE) and k-nearest neighbors (KNN) are the best for 100 data points at high and low noise-to-signal levels, respectively, whereas KNN is the best for 1000 data points consistently across noise levels. One real application of nonlinear dependence based on MI is to capture extrabasinal connections between El Niño-Southern Oscillation (ENSO) and river flows in the tropics and subtropics, specifically the Nile, Amazon, Congo, Paraná, and Ganges rivers which reveals 20-70% higher dependence than those suggested so far by linear correlations. For extremes analysis, this dissertation develops a new measure *precipitation extremes volatility index* (PEVI), which measures the variability of extremes, is defined as the ratio of return levels. Spatio-temporal variability of PEVI, based on the Poisson-generalized Pareto (Poisson-GP) model, is investigated on weekly maxima observations available at 2.5^0 grids for 1940-2004 in South America. From 1965-2004, the PEVI shows increasing trends in few parts of the Amazon basin and the Brazilian

highlands, north-west Venezuela including Caracas, north Argentina, Uruguay, Rio De Janeiro, São Paulo, Asuncion, and Cayenne. Catingas, few parts of the Brazilian highlands, São Paulo and Cayenne experience increasing number of consecutive 2- and 3-days extremes from 1965-2004. This dissertation also addresses the ability to detect the chaotic signal from a finite time series observation of hydrologic systems. Tests with simulated data demonstrate the presence of thresholds, in terms of noise to chaotic-signal and seasonality to chaotic-signal ratios, beyond which the set of currently available tools is not able to detect the chaotic component. Our results indicate that the decomposition of a simulated time series into the corresponding random, seasonal and chaotic components is possible from finite data. Real streamflow data from the Arkansas and Colorado rivers do not exhibit chaos. While a chaotic component can be extracted from the Arkansas data, such a component is either not present or can not be extracted from the Colorado data.

Chapter 1

Introduction

In hydrology and climate, the presence of nonlinear dependence and chaos has strong implications for predictive modeling and the analysis of dominant processes. Analysis of hydrological and climatological extremes may also aid in predictive modeling by giving enhanced understanding of processes driving the extremes and perhaps delineate possible anthropogenic or natural causes. This dissertation analyzes and develops three components of predictive modeling, specifically nonlinear dependence, extremes, and chaos, and tests their viability and scalability on the real data. This section introduces these components, describes the state of the art in each component, points gaps in their respective literatures, and outlines procedures for targeting their respective gaps.

In nonlinear systems, the understanding of underlying nonlinear processes and their interactions are very important for predictive modeling as well as for generating bounds on predictability. However, data analysis methods based on nonlinear dynamical approaches are typically not robust when applied to short and noisy data [1]. The definition of what constitutes short and noisy, in terms of data sizes and noise-to-signal ratios, may be application and context specific. A consideration of data availability scenarios in a couple of domains, specifically the earth sciences and biomedical engineering, in conjunction with the literature on mutual information (MI) estimation methods, suggest that a critical gap continues to exist in our understanding of situations where the length of data sets is short, particularly of the order of 100 or 1 000 data points. Linear correlation may not be an adequate measure of dependence even for simple nonlinear functional forms. This can be simply shown in the case of two variables (X, Y), where $(Y = X^2)$, and X is uniformly distributed in the interval $(-1, 1)$. The theoretical covariance and hence the linear correlation reduces to zero even though the variable Y is completely specified once X is known. The situation gets even more problematic when the nonlinear interactions get more complex. The problem of detecting excessive spurious dependence or missing existing dependence structures among nonlinear signals is exacerbated for short and noisy data. The degree to which even small amount of noise can obscure the underlying dependence structure is evident from Fig. 1 which shows two cases, such as quadratic and Henon, based on simulations with 100 points each. In both cases, the simulated data are contaminated with Gaussian noise with zero mean and standard deviation given

by σ_ϵ/σ_s, which is called the noise-to-signal ratio. The variables σ_ϵ and σ_s are the standard deviations of the noise and signal, respectively. Visual inspection reveals that the dependence structure departs significantly from the underlying true dependence structure as the noise-to-signal ratio increases. Robust measures for nonlinear dependence would need to capture the dependence structure even when the latter is obscured by noise. Several methods for the estimation of MI-based nonlinear dependence have been suggested in recent years, such as kernel denstiy estimators (KDE) [2], adaptive partitioning of the observation space [3], Parzen window density estimator [4], k-nearest neighbors (KNN) [5], Edgeworth approximation of differential entropy (Edgeworth) [6], mutual information carried by the rank sequences [7], and adaptive partitioning of the XY plane (referred here as Cellucci) [8]. Previous studies [2, 5, 8, 9] designed to compare existing or newly proposed methods for nonlinear dependence with each other, as well as with a standard method, have been limited in scope. The goal of this dissertation is to investigate and compare recently developed MI estimation methods, specifically KDE, KNN, Edgeworth, and Cellucci, based on simulated data generated from linear, quadratic, periodic, and chaotic data contaminated artificially with various levels of Gaussian noise. The performance of the MI-estimation methods are compared against each other and against baselines comprising linear correlation coefficient (CC) obtained from linear regression (LR) and rank-based CCs from Kendall's τ [10]. We have also used theoretical MI values from linear, quadratic, and periodic, which can be computed analytically, for comparing the performance of different MI-estimation methods. The purpose of the above comparisons is to identify the one MI-estimation method or combination of MI-estimation methods in terms of robustness to short and noisy data, at least for the illustrations considered here, whose estimation values are closest to the theoretical MI values and significantly different from linear estimates in that their confidence bounds do not intersect.

ENSO events impact regional precipitation in the tropics and subtropics, ultimately causing inter-annual variability in river flows. The ocean-atmosphere-land interactions are complex and far from being completely understood and accurately modeled. A slight disturbance in these interactions would usually result in sometimes surprising distant correlations and climate patterns. Analyses of the rainfall anomalies during the warm (El Niño) and cold (La Niña) episodes of ENSO suggest the existence of nonlinear sea surface temperature (SST)-rainfall relationships in the tropics and a strong influence of SST forcing on equatorial rainfall in the geographic vicinity of that forcing [11]. To properly explain and ultimately predict this variability, it is important to disentangle, as far as possible, long range climatic phenomena from recent effects such as those possibly produced by deforestation and global warming. While the relationships among many climate and hydrological variables are decidedly nonlinear [12], *linear* dependence measures are still being used as a

matter of course to relate ENSO and inter-annual variability in river flows. These measures have ranged from linear correlation coefficients (CC) in the time domain [13–16] to the cross-spectrum analysis [17, 18]. One of the reasons for using linear measures is that the inherent noise and periodicity in the observations together with short length of the available sample sizes make it difficult to use nonlinear approaches in climate and hydrology [19–21]. The goal of this dissertation is to investigate the nonlinear dependence between ENSO and the annual flow of some of the largest tropical and subtropical rivers, specifically the Nile, Amazon, Congo, Paraná and Ganges, through a MI-based measure.

Precipitation extremes can have significant impacts on human society, economics, and nature. An understanding of the intensity and frequency of precipitation extremes can be very useful for infrastructure development to prevent flooding and landslides, as well as for water resources and agricultural management. A better understanding of precipitation extremes can help hydrologic scientists and climatologists gain enhanced understanding of precipitation processes driving the extremes and perhaps delineate possible anthropogenic or natural causes. Previous studies investigated trends and variability of precipitation extremes in many parts of the world in the twentieth century, specifically the United States [22, 23], India [24], Southeast Asia and the South Pacific [25], Australia [23, 26], Europe [27], Caribbean [28], Italy [29], Balkans [30], Canada, Norway, Russia, China, Mexico [23], Japan [31], Sweden [32], southeastern South America [33], and the state of São Paulo, Brazil [34]. Recently the spatio-temporal variability of dependence among precipitation extremes was investigated over the entire South America for the period 1940-2004 using a new approach (suggested by Kuhn [35]) [36]. However, we are not aware of any prior investigations on spatial and temporal variability of precipitation extremes over the entire continent of South America. The generalized extreme value (GEV) distribution, developed by Jenkinson [37], has been traditionally utilized for modeling precipitation extremes [38–40]. This approach is also called the block maxima approach since it fits the distribution to the highest values in blocks of equal size, e.g., maximum yearly precipitation. It has some advantages, e.g., its requirements can be met by a simplified summary of data and the block maxima can be assumed to be independent random variables [41]. But the main drawback of the GEV distribution is that it does not utilize all the available information about the upper tail of the distribution, e.g., two highest extreme precipitation events may occur in the same year [41]. An alternative approach is to use peaks over threshold (POT) which was originated in hydrology and makes use of all the data available, e.g., all daily precipitation data [42]. The statistical model underlying the POT method consists of (1) Poisson process for the occurrences of extremes over a large threshold and (2) generalized Pareto (GP) distribution (with scale (σ) and shape (ξ) parameters), developed by Pickands [43], for the distribution of excesses over a large threshold. This model is also termed

as Poisson-GP model. Recently, the GP distribution has been utilized for modeling threshold excesses from daily precipitation data [44,45]. This dissertation utilizes the Poisson-GP model for investigating the spatial and temporal variability of precipitation extremes using daily precipitation data in 2.5^0 grids South America for 1940-2004 [46]. The Poisson-GP model assumes the data to be independent and identically distributed (IID) [39]. A long-term trend and seasonality in the data violate the assumption of identically distributed data whereas the assumption of independent data is violated if there is temporal dependence in the data [47]. In order to check the IID assumption for the Poisson-GP model, we consider three different sets of data based on this daily data: daily data itself, weekly maxima, and weekly maxima residuals. Weekly maxima residuals are obtained by subtracting the long term mean of weekly maxima of a particular week, i.e., mean of maximum weekly precipitation across the same week for all years used in the analysis, from weekly maxima of the same week. These datasets are compared to choose the best data by examining temporal dependence through auto-correlations and seasonal trends. In order to check the quality of the Poisson-GP model, we also compare these datasets in terms of the Poisson property of the occurrences of extremes and quality of the GP distribution. The scale of the data and the need for efficient computations, which can be eventually automated, preclude choosing thresholds based on human judgment. We choose thresholds as 95%-quantile for weekly maxima and weekly maxima residuals and 99%-quantile for the daily data at each grid point. Spatial variability is investigated for 65 years (1940-2004) and the last 40 years (1965-2004) are also studied for the temporal variability with 25-year moving window, i.e., 1965-1989, 1966-1990,. . ., 1980-2004. The temporal variability is given by the slope of a linear trend obtained by fitting a regression line to 16 values from 16 time windows from 1965-2004. We investigate the spatial and temporal variability of thresholds, two parameters of the GP distribution (σ and ξ) and their standard errors, 50-year and 200-year return levels (RL), and *precipitation extremes volatility index* (PEVI) [46]. The PEVI measures the variability of extremes and is defined as a ratio of RLs. Based on daily data, the spatial and temporal variability of the number of consecutive 2-days and 3-days extremes and the spatial variations of the number of monthly extremes are investigated [46].

The presence of chaos in hydrology has been suggested by previous researchers [48–64]. The ability to detect and model chaotic behavior from finite hydrologic time series has recently been debated [65, 66]. Characterization of chaos from real-world observations is known to be a difficult problem in nonlinear dynamics [67–69]. The complexity was highlighted in the context of climate models by [70], who demonstrated that sensitivity to initial conditions may become less apparent when the randomness in internal atmospheric variables begins to dominate. Fundamental questions still remain unanswered in these areas, for example

the ability to detect chaos from a finite time series with random and seasonal components, the ability to decompose a time series into these components, and the corresponding implications for predictive modeling. However, addressing these questions is critical for hydrology. This can be gauged from the wealth of hydrologic literature in areas like complexity analysis [71–73], predictability [64] and nonlinear predictive modeling [74–77]. This dissertation investigates the ability of nonlinear dynamical tools to detect, characterize, and predict chaos from finite hydrologic observations, using both simulated and real time series. Realistic simulated data is generated by contaminating chaotic signals with random and seasonal components, while real streamflow data are used from the Arkansas and Colorado rivers. The correlation dimension method is used for detecting the possible presence of chaos. Nonlinear predictive models, namely the phase-space reconstruction (PSR) and artificial neural networks (ANN) are employed for time series decomposition and prediction. The presence of thresholds for the detect-ability of chaos is demonstrated, specifically when a chaotic signal is mixed with random or seasonal signals, or a combination thereof. These thresholds can be expressed in terms of the relative dominance of the chaotic component compared to the random or seasonal components. The ability to decompose a time series into the contributions from the individual components (random, seasonal and chaotic) is shown.

Chapter 2

Motivation

In geophysics, the ability to predict may be increased (a) by including nonlinear relationships between geophysical processes, (b) by generating t-year return levels (levels expected to be exceeded on average once every t years) based on the analysis of low probability and high risk geophysical events (also called extremes), and (c) if there is a presence of chaos, which implies short-term predictability, in geophysical processes. Since the geophysical data are generally short and noisy, there is a need to develop and investigate nonlinear dependence, extremes, and chaos detection measures, which are robust to finite and noisy data. Mutual information has been frequently utilized for capturing the complete dependence structure including nonlinear dependence. Recently, several methods have been proposed for the MI estimation, such as KDE [2], KNN [5], Edgeworth approximation of differential entropy [6], and adaptive partitioning of the XY plane [8]. However, outstanding gaps in the current literature have precluded the ability to effectively automate these methods, which, in turn, have caused limited adoptions by the application communities. This dissertation attempts to address a key gap in the literature, specifically, the evaluation of the above methods to choose the best method, particularly in terms of their robustness for short and noisy data, based on comparisons with the theoretical MI estimates, which can be computed analytically, as well with linear correlation and Kendall's τ. In addition, there is also a need to find an optimal smoothing parameter for a Gaussian kernel for KDE and optimal number of nearest neighbors for KNN when the data are short and noisy. An understanding of the intensity and frequency of precipitation extremes can be very useful for infrastructure development to prevent flooding and landslides, as well as for water resources and agricultural management. Previous studies [22–32] investigated trends and variability of precipitation extremes in many parts of the world in the twentieth century but no study was found focussing on precipitation extremes over the entire continent of South America. Since the nations of South America are developing, highly populated, and not capable enough to respond to disasters caused by precipitation extremes, there is a clear need to investigate spatial and temporal variability of precipitation extremes in South America. Extreme value distributions generate information about extremes in terms of their parameters which can be understood only by statisticians. There is a need to develop a measure which should be statistically valid, easily quantified and visualized over large geographical areas, and understood

not only by statisticians but also by hydrologists, climatologists, and decision-makers. In hydrological data, the presence of noise and seasonality makes the chaos detection process challenging. There is a need to define thresholds, in terms of noise to chaotic-signal and seasonality to chaotic-signal ratios, beyond which the set of currently available tools is not able to detect the chaotic component. If there is chaos in the data, it would be very interesting to investigate if it is possible to separate chaotic and random components from finite data. Previous studies [52, 53, 55, 56] investigated the presence of chaos in many rivers around the world but no study was found focussing on two major rivers, i.e., Arkansas and Colorado rivers, in the United States. This dissertation targets three components, i.e., nonlinear dependence, extremes, and chaos, of predictive modeling separately but there is also a need to look at the possibility of inter-connecting these components for improving predictive models.

Chapter 3

Relative Performance of Mutual Information Estimation Methods for Quantifying the Dependence Among Short and Noisy Data

Commonly used dependence measures, such as linear correlation, cross-correlogram or Kendall's τ, cannot capture the complete dependence structure in data unless the structure is restricted to linear, periodic or monotonic. Mutual information (MI) has been frequently utilized for capturing the complete dependence structure including nonlinear dependence. Recently, several methods have been proposed for the MI estimation, such as kernel density estimators (KDE), k-nearest neighbors (KNN), Edgeworth approximation of differential entropy, and adaptive partitioning of the XY plane. However, outstanding gaps in the current literature have precluded the ability to effectively automate these methods, which, in turn, have caused limited adoptions by the application communities. This dissertation attempts to address a key gap in the literature, specifically, the evaluation of the above methods to choose the best method, particularly in terms of their robustness for short and noisy data, based on comparisons with the theoretical MI estimates, which can be computed analytically, as well with linear correlation and Kendall's τ. Here we consider smaller data sizes, such as 50, 100, and 1 000, where this dissertation considers 50 and 100 data points as *very short* and 1 000 as *short*. We consider a broader class of functions, specifically linear, quadratic, periodic and chaotic, contaminated with artificial noise with varying noise-to-signal ratios. Our results indicate KDE as the best choice for *very short* data at relatively high noise-to-signal levels whereas the performance of KNN is the best for *very short* data at relatively low noise levels as well as for *short* data consistently across noise levels. In addition, the optimal smoothing parameter of a Gaussian kernel appears to be the best choice for KDE while three nearest neighbors appear optimal for KNN. Thus, in situations where the approximate data sizes are known in advance, and exploratory data analysis and/or domain knowledge can be used to provide *a priori* insights on the noise-to-signal ratios, the results in the paper point to a way forward for automating the process of MI estimation.

3.1 Introduction

In nonlinear systems, the understanding of underlying nonlinear processes and their interactions are very important for predictive modeling as well as for generating bounds on predictability. However, data analysis methods based on nonlinear dynamical approaches are typically not robust when applied to short and noisy data [1]. The definition of what constitutes short and noisy, in terms of data sizes and noise-to-signal ratios, may be application and context specific. A consideration of data availability scenarios in a couple of domains, specifically the earth sciences and biomedical engineering, in conjunction with the literature on mutual information (MI) estimation methods, suggest that a critical gap continues to exist in our understanding of situations where the length of data sets is short, particularly of the order of 100 or 1 000 data points.

Physically-based definitions for what constitutes *long* versus *short* data sizes need to follow from a comparison of sampling coverage time-span vis-$\grave{a}$-vis the characteristic period of the dynamical system under consideration. The characteristic period can be, for example, one full seasonal cycle for purely seasonal observations, or a complete span of the attractor for a chaotic system. If the sample size is large but the sampling coverage is restricted to a small portion of the cycle or the attractor, then observations are still not representative of the population. In this sense, the data size must still be considered *short* in a physical sense because they do not have the coverage necessary to make the relevant inferences from the data. While samples with greater coverage is more representative of the population, the tradeoff, especially for a limited number of samples, is that the sampling frequency needs to be adequate to capture the features of the dynamical system and make appropriate inferences from the observations. The sampling frequency in this sense is related to the Nyquist frequency of the system. In this sense, even if the sampling coverage is large but the frequency is inadequate, the data size must still be considered *short* from a physical perspective. Thus, the Nyquist frequency on the one hand, and the characteristic period of the dynamical system under consideration on the other, provide guidelines for the definitions of *long* versus *short* data sizes, and indeed provides a physical basis for such definitions. However, in real-world situations, the knowledge of the characteristic period of the dynamical system or the signal bandwidth may not necessarily be known *a priori*, and in some cases, may be difficult to estimate if the data are contaminated with non-repeatable patterns, measurement errors, or other forms of noise. Thus, for such systems, there is a need for caution before making a claim that a set of observations is *short* or, perhaps more important, *long* enough. This paper is concerned with simulated data, where we have knowledge of the system, and generates noise sequences from independent and identically distributed processes. Here we implicitly define the characteristic time (basic period) of the system as equal

to unity, thus the number of data points is a natural measure for our examples. In this dissertation, a data size of 50 to 100 is referred as *very short* whereas a data size of 1 000 is considered *short*.

We use the term noise in a generic sense to include variability in measurement errors as well as any inherent, but non-repeatable, randomness that may be present in complex systems. Indeed, noise levels encountered in real-world data may vary considerably depending on the domain, data collection methods, measurement accuracy, inherent randomness in the observables, as well as other factors. Here we consider noise-to-signal ratios that range all the way from zero, which implies no noise, to unity, which implies that the noise is as dominant as the underlying signal itself. For this dissertation, we call a noise-to-signal ratio of zero to about a half as *low noise* and higher ratios as *high noise*.

Linear correlation may not be an adequate measure of dependence even for simple nonlinear functional forms. This can be simply shown in the case of two variables (X, Y), where $(Y = X^2)$, and X is uniformly distributed in the interval $(-1, 1)$. The theoretical covariance and hence the linear correlation reduces to zero even though the variable Y is completely specified once X is known. The situation gets even more problematic when the nonlinear interactions get more complex. However, the application of nonlinear dynamical and/or information theoretic measures of dependence can be a challenge, especially when short and noisy data are available. For example, the identification of the underlying nonlinear dynamical component via the correlation dimension is known to be difficult problem for geophysical [21] or electroencephalographic (EEG) [78,79] signals. Similarly, the detection of the underlying interactions among variables characterizing a complex system becomes a difficult task [80]. The inherent difficulty of numerical estimation as well as perceived problems with model parsimony or overfitting have resulted in relatively limited use of nonlinear approaches, even when the underlying processes are known to be nonlinear. The problem exists in certain biomedical applications [81–83], but grows more acute in domains like geophysics [21, 84] where the data collection and generation processes are often not repeatable. Our definition of what constitutes short and noisy data is motivated from problems in these domains. The references cited earlier show that *very short* and *short* data sets, as well as *low noise* and *high noise* conditions do exist for real-world problems. Thus, there is a clear need to investigate methods, which are robust to short and noisy data, for the determination of nonlinear multivariate interactions. However, the methodologies need to be rigorously tested such that well-known problems in nonlinear statistics like overfitting do not yield misleading correlations.

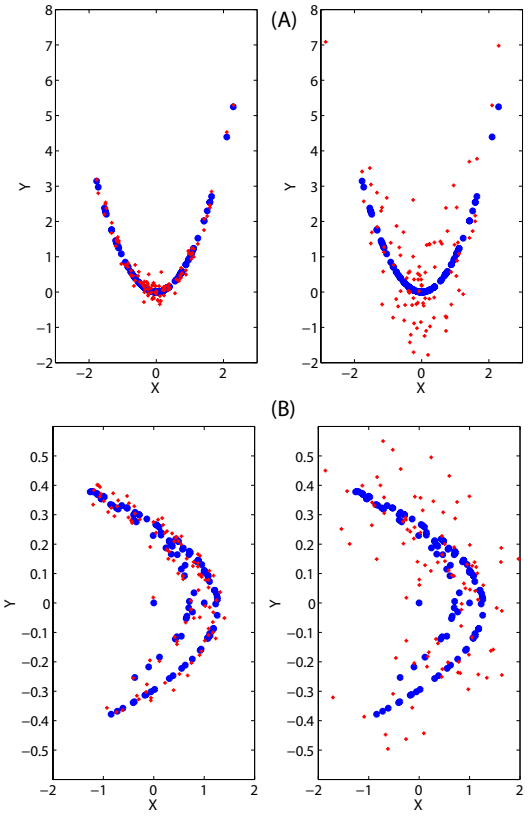

Figure 1. Plot of 100 points with different noise-to-signal ratios (shown by plus) and with zero noise level (shown by dots). Noise-to-signal ratios on the left and right figures are 0.1 and 0.5, respectively. (A) $X \sim N(0,1),\ Y : y_i = {x_i}^2 + \varepsilon_i$, where $\varepsilon \sim N(0, \sigma_\varepsilon)$ is the Gaussian noise with zero mean and σ_ε standard deviation. (B) $X : x_i = H_{x_i} + \varepsilon x_i,\ Y : y_i = H_{y_i} + \varepsilon y_i$, where H_X and H_Y are the X and Y components of the Henon map, respectively. $\varepsilon x \sim N(0, \sigma_{H_X})$ and $\varepsilon y \sim N(0, \sigma_{H_Y})$, where σ_{H_X} and σ_{H_Y} are the standard deviations of H_X and H_Y, respectively.

The problem of detecting excessive spurious dependence or missing existing dependence structures among nonlinear signals is exacerbated for short and noisy data. The degree to which even small amount of noise can obscure the underlying dependence structure is evident from Fig. 1 which shows two cases, such as quadratic and Henon, based on simulations with 100 points each. In both cases, the simulated data are contaminated with Gaussian noise with zero mean and standard deviation given by σ_ϵ/σ_s, which is called the noise-to-signal ratio. The variables σ_ϵ and σ_s are the standard deviations of the noise and signal, respectively. Visual inspection reveals that the dependence structure departs significantly from the underlying true dependence structure as the noise-to-signal ratio increases. Robust measures for nonlinear dependence would need to capture the dependence structure even when the latter is obscured by noise. Previous studies designed to compare existing or newly proposed methods for nonlinear dependence with each other, as well as with a standard method, have been limited in scope. The classic algorithm was proposed by Fraser and Swinney [9], which was compared with the kernel density estimation (KDE) method given by Moon et al. [2]. The comparison utilized the following combination of data sizes and simulations: 400 for a sinusoidal curve, 500 for an autoregressive process, 4 096 for data sets generated from the Lorenz system, and 2 048 for Rossler, where the last two are chaotic. Later, KDE was refined and validated on real-world geophysical data sets [85]. Kraskov et al. [5] compared two k-nearest neighbors (KNN) estimators with simulations from correlated Gaussians for data sizes of 125, 250, 500, 1 000, 2 000, 4 000, 10 000 and 20 000, as well as with simulations from the exponential distribution. In addition, they tested their methods on gene expression data. The Edgeworth approximation of differential entropy proposed by Hulle [6] was compared against the KNN method and Parzen density estimator. For the comparisons, the data sets of size 1 000 and 10 000 were generated from the Gaussian and exponential distributions. Cellucci et al. [8] focused on statistical evaluation of mutual information estimation by comparing the MI estimates with linear correlations and the rank-based correlations from Kendall's τ. In addition, they proposed a new algorithm based on adaptive partitioning and compared it with the Fraser-Swinney method given in [9]. Their comparisons utilized simulations from the Gaussian distribution, linear and quadratic functions contaminated with artificial noise, as well as the chaotic systems, such as Lorenz and Rossler. The data sizes utilized for the comparison were 4 096, 8 192, 10 000, 65 536 and 100 000. Cellucci et al. [8] mentioned that when they initiated their research, the KNN method by Kraskov et al. [5] had not been published yet. Indeed, they also suggested a need of an expanded future research effort to compare and contrast their adaptive partitioning method with the KNN and KDE methods. From these discussions, it is clear that a thorough comparison of the various methods for the estimation of MI, specifically, KDE, KNN, Edgeworth and adaptive partitioning, do not exist in the literature. Furthermore, detailed

comparisons have not been attempted across a wide class of simulated functional forms. In addition, the MI estimation methods have not been compared with baseline approaches like linear correlation and Kendall's τ, other than the specific comparisons presented by Cellucci et al. [8]. Finally, a clear gap exists in terms of detailed comparisons of the various MI estimation methods for short and noisy data.

As discussed earlier, several methods for the estimation of MI have been suggested in recent years, such as KDE [2], adaptive partitioning of the observation space [3], Parzen window density estimator [4], KNN [5], Edgeworth approximation of differential entropy (Edgeworth) [6], mutual information carried by the rank sequences [7], and adaptive partitioning of the XY plane (referred here as Cellucci) [8]. The goal of this dissertation is to investigate and compare recently developed MI estimation methods, specifically KDE, KNN, Edgeworth, and Cellucci, based on simulated data generated from linear, quadratic, periodic, and chaotic data contaminated artificially with various levels of Gaussian noise. We generate 50, 100, and 1 000 points for our analysis. As mentioned earlier, the motivation for the data sizes comes from a specific geophysical application (the relationship of the interannual climate index known as ENSO with the variability of tropical riverflows [84]) and a specific biomedical application (dependence among EEG signals [81, 82]). The simulated data allow us to compare the relative performance of the MI estimation methods across an order of magnitude in terms of data sizes and noise-to-signal ratios ranging from 0 to 1 in increments of 0.1. Uncertainties on the MI estimates are obtained through bootstrapping and provided as 90% confidence bounds. The total number of bootstraps used for 50, 100, and 1 000 points are 200, 100, and 10, respectively, reflecting a pragmatic trade-off between the need for accuracy and computational tractability. However, such trade-offs may not be required in more efficient or higher performance computational implementations. The performance of the MI-estimation methods are compared against each other and against baselines comprising linear correlation coefficient (CC) obtained from linear regression (LR) and rank-based CCs from Kendall's τ. We have also used theoretical MI values from linear, quadratic, and periodic, which can be computed analytically, for comparing the performance of different MI-estimation methods. The purpose of the above comparisons is to identify the one MI-estimation method or combination of MI-estimation methods in terms of robustness to short and noisy data, at least for the illustrations considered here, whose estimation values are closest to the theoretical MI values and significantly different from linear estimates in that their confidence bounds do not intersect.

The rest of the paper is organized as follows. In Sec. 3.2, the MI and its estimation methods are described. The MI is defined in Sec. 3.2A while we outline the four MI estimation methods, namely KDE, KNN, Edgeworth, and Cellucci, in Sec. 3.2B. In Sec. 3.3, the description of simulated data sets to be analyzed is

provided. We present and discuss the results obtained using four MI estimation methods, LR, and Kendall's τ in Sec. 3.4. In Sec. 3.5, the conclusion and discussion are presented.

3.2 Mutual information and its estimation methods

Several dependence measures, such as linear correlation, cross-correlogram, Kendall's τ, and MI, have been utilized to capture the dependence structure between a pair of variables (X, Y). However, while the first three measures can only capture linear, periodic or monotonic dependence, MI can describe the full dependence structure including nonlinear dependence if any [86]. In addition, MI reduces to the linear dependence when the data are indeed linearly related. In an information theoretic sense, MI quantifies the information stored in one variable about another variable. MI has several satisfying theoretical properties and analogous relations with the linear correlation. While the linear CC can be used to calculate the prediction mean squared errors (MSE) from linear regression, MI can be used to compute a bound on the achievable prediction MSE based on the information content in the independent variables about the dependent variables. MI has been shown to have traditional analysis of variance (ANOVA)-like interpretations [87]. For time serial data, MI can be computed as a function of temporal lags to obtain nonlinear versions of the auto- or cross-correlation (ACF or CCF) functions. The information theoretic properties of MI, which make it a reliable measure of statistical dependence, have been described by Cover and Thomas [88]. The applicability of MI for feature, parameter and model selection problems have been described by Brillinger [87]. Besides the direct use of MI in the computation of nonlinear dependence [87, 89], MI has indicated value in areas ranging from optimal time delay embeddings during phase-space reconstructions [9] to extracting causal relationships among variables ([90,91]) [8].

3.2.1 Definitions of mutual information

For the bivariate random variables (X, Y), the MI is defined as

$$I(X;Y) = \int_Y \int_X p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x) p_Y(y)} \, dx \, dy, \tag{1}$$

where $p_{XY}(x, y)$ is the joint probability density function (*pdf*) between X and Y; and $p_X(x)$ and $p_Y(y)$ are the marginal *pdfs* [8]. The unit of MI is defined corresponding to the base of the logarithm in Eq. (1), i.e., nats for log, bits for $\log_2$, and Hartleys for $\log_{10}$. MI is positive and symmetrical, i.e., $I(X;Y) = I(Y;X)$. It is also invariant under one to one transformations, i.e., $I(X;Y) = I(U;V)$, where $u = f(x)$, $v = f(y)$,

and f is invertible. If X and Y are independent, the joint *pdf* is equal to the product of marginal *pdfs* leading to $I(X;Y) = 0$ from Eq. (1). If there exists perfect dependence between X and Y, MI approaches infinity.

MI between random variables X and Y can also be defined in terms of information entropies as

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y), \end{aligned} \tag{2}$$

where $H(X)$ and $H(Y)$ are called the marginal information entropies which measure the information content in X and Y, respectively, $H(Y|X)$ is the entropy of Y conditional on X which measures the information content remaining in Y if the information content in X is known completely, and $H(X,Y)$ is the joint information entropy which measures the information content in a pair of random variables X and Y. The bivariate case is considered here for simplicity.

The linear CC (ρ) between two variables X and Y is a measure of the strength of the linear dependence between the variables and varies from 0 to 1. The estimation of the most likely value and the corresponding uncertainties are relatively straightforward. However, the estimation of the mean and uncertainty bounds, for an MI-based dependence measure that is normalized to scale between 0 to 1, is an area of ongoing research.

If (X,Y) is bivariate normal, the MI and linear CC are related as $I(X;Y) = -0.5\log[1-\rho(X,Y)^2]$ [92]. Joe [93] proposed a linear CC like measure for MI, which scales from 0 to 1, given as

$$\hat{\lambda}(X,Y) = \sqrt{1 - \exp[-2\hat{I}(X;Y)]}, \tag{3}$$

where $\hat{\lambda}(X,Y)$ and $\hat{I}(X;Y)$ are the estimated nonlinear CC and MI, respectively. Later Granger and Lin [94] used the same measure to estimate nonlinear CC from the MI. While this dissertation utilizes nonlinear CC based solely on MI, other bases for nonlinear CC suggested in the literature include mutual nonlinear prediction [95] and nonlinear association analysis [96]. A detailed comparison of the various definitions of nonlinear CC and their relative performances are left as areas for future research. In order to estimate the predictability of Y given X, once the MI is known, Brillinger [87] proposed an equation which provides a lower bound on the prediction MSE. This equation, which is analogous to the MSE for linear regression obtained from the linear correlation coefficient, is given as

$$\widehat{MSE}(Y) \geq \frac{1}{2\pi e}\exp[2\{\hat{H}(Y) - \hat{I}(X;Y)\}],$$

where $\hat{H}(Y)$ is the estimated information entropy of Y and $\widehat{MSE}(Y)$ gives a lower bound on MSEs from the MI and measures the predictability of Y based on the information content in X.

3.2.2 Mutual information estimators

3.2.2.1 Kernel density estimators (KDE) The MI in Eq. (1) for any bivariate data set *(X, Y)* of size n can be estimated as

$$\hat{I}(X;Y) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\hat{p}_{XY}(x_i, y_i)}{\hat{p}_X(x_i)\hat{p}_Y(y_i)}\,, \tag{4}$$

where $\hat{p}_{XY}(x_i, y_i)$ is the estimated joint *pdf* and $\hat{p}_X(x_i)$ and $\hat{p}_Y(y_i)$ are the estimated marginal *pdfs* at (x_i, y_i).

For the multivariate data set $(\boldsymbol{x_1}, .., \boldsymbol{x_n})$, where each $\boldsymbol{x}$ is in a d-dimensional space, the multivariate kernel density estimator with kernel K is defined by

$$\hat{p}(\boldsymbol{x}) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\boldsymbol{x} - \boldsymbol{x_i}}{h}\right), \tag{5}$$

where h is the smoothing parameter [97]. We choose the standard multivariate normal kernel defined by

$$K(\boldsymbol{x}) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right). \tag{6}$$

Using Eqs. (5) and (6), the probability density function is defined as

$$\hat{p}(\boldsymbol{x}) = \frac{1}{nh^d}\sum_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^d|\mathbf{S}|}} \exp\left(-\frac{(\boldsymbol{x} - \boldsymbol{x}_i)^T\mathbf{S}^{-1}(\boldsymbol{x} - \boldsymbol{x}_i)}{2h^2}\right), \tag{7}$$

where $\mathbf{S}$ is the covariance matrix and $|\mathbf{S}|$ is the determinant of $\mathbf{S}$. For a normal kernel, Silverman [97] suggested an optimal smoothing parameter or Gaussian bandwidth given as

$$h_o = \left(\frac{4}{d+2}\right)^{1/(d+4)} n^{-1/(d+4)}. \tag{8}$$

Moon et al. [2] presented the same procedure and utilized Eq. (7) for estimating marginal probability densities, i.e., $\hat{p}_X$ and $\hat{p}_Y$, and the joint probability density, i.e., $\hat{p}_{XY}$, and substituted these densities in Eq. (4) to estimate MI.

3.2.2.2 k-nearest neighbors (KNN) If $X = (\boldsymbol{x_1}, .., \boldsymbol{x_n})$, where each $\boldsymbol{x}$ is in d-dimensional space, is a continuous random variable, the Shannon entropy of X, defined as

$$H(X) = -\int p(\boldsymbol{x}) \log p(\boldsymbol{x})\, d\boldsymbol{x},$$

can be estimated by

$$\hat{H}(X) = -\frac{1}{n}\sum_{i=1}^{n} \log \hat{p}(\boldsymbol{x}_i), \tag{9}$$

where $\hat{p}(\boldsymbol{x}_i)$ is the estimated marginal *pdf* at $\boldsymbol{x}_i$. Kraskov et al. [5] expanded Eq. (9) as

$$\begin{aligned} \hat{H}(X) = -\frac{1}{n}\sum_{i=1}^{n}\psi(n_x(i)) \quad &- \quad \frac{1}{k} + \psi(n) + \log c_{d_X} \\ &+ \quad \frac{d_X}{n}\sum_{i=1}^{n} \log \epsilon(i), \end{aligned} \tag{10}$$

where n and k are the number of data points and nearest neighbors, respectively; d_X is the dimension of $\boldsymbol{x}$; and c_{d_X} is the volume of the d_X-dimensional unit ball. For two random variables X and Y, let $\epsilon(i)/2$ be the distance between (x_i, y_i) and its kth neighbor denoted by (kx_i, ky_i). Let $\epsilon_x(i)/2$ and $\epsilon_y(i)/2$ be defined as $||x_i - kx_i||$ and $||y_i - ky_i||$, respectively. $n_x(i)$ is the number of points x_j such that $||x_i - x_j|| \leq \epsilon_x(i)/2$. $\psi(x)$ is the digamma function, $\psi(x) = \Gamma(x)^{-1} d\Gamma(x)/dx$, which satisfies the relation $\psi(x+1) = \psi(x) + 1/x$, with $\psi(1) = -C$, where $C = 0.5772156649$ is the Euler-Mascheroni constant. Similarly, $\hat{H}(Y)$ can be derived by replacing x with y in Eq. (10). In the similar way, the estimated joint entropy between X and Y can be given as

$$\begin{aligned} \hat{H}(X,Y) = -\psi(k) \quad &- \quad \frac{1}{k} + \psi(n) + \log\left(c_{d_X} c_{d_Y}\right) \\ &+ \quad \frac{d_X + d_Y}{n}\sum_{i=1}^{n} \log \epsilon(i), \end{aligned}$$

where d_Y is the dimension of $\boldsymbol{y}$; and c_{d_Y} is the volume of the d_Y-dimensional unit ball. Substituting $\hat{H}(X)$, $\hat{H}(Y)$, and $\hat{H}(X,Y)$ in Eq. (2), the MI can be estimated as

$$\hat{I}(X;Y) = \psi(k) - \frac{1}{k} - \frac{1}{n}\sum_{i=1}^{n}[\psi(n_x(i)) + \psi(n_y(i))] + \psi(n),$$

where $n_y(i)$ is the number of points y_j such that $||y_i - y_j|| \leq \epsilon_y(i)/2$ [5].

3.2.2.3 Edgeworth approximation of differential entropy (Edgeworth) If $X = (\boldsymbol{x_1}, .., \boldsymbol{x_n})$, where each $\boldsymbol{x}$ is in a d-dimensional space, the Edgeworth expansion of the density $p(\boldsymbol{x})$ after ignoring higher order terms is given by

$$p(\boldsymbol{x}) \approx \phi_p(\boldsymbol{x}) \left(1 + \frac{1}{3!}\sum_{i,j,k} \kappa^{i,j,k} h_{i,j,k}(\boldsymbol{x})\right), \tag{11}$$

where $\phi_p(\boldsymbol{x})$ is the normal distribution with the same mean and covariance matrix as p; (i, j, k) is the input dimension where $(i, j, k) \in (1, \ldots, d)$; $\kappa^{i,j,k}$ is the standardized cumulant, i.e., $\kappa^{i,j,k} = \frac{\kappa^{ijk}}{\sigma_i \sigma_j \sigma_k}$, where κ^{ijk} is the cumulant for input dimensions (i, j, k) and σ is the standard deviation, for large number of points; and $h_{i,j,k}$ is the ijkth Hermite polynomial [6].

Let $p(\boldsymbol{x})$ be defined in a set $\mathbb{X}$. The differential entropy of X is defined as

$$H(X) = -\int_{\mathbb{X}} p(\boldsymbol{x})\ \log p(\boldsymbol{x})\ d\boldsymbol{x}.$$

In terms of the density, i.e., $p(\boldsymbol{x})$, defined in Eq. (11), the differential entropy of X can also be defined as

$$\begin{aligned} H(p) &= H(\phi_p) - J(p) \\ &= H(\phi_p) - \int_{\mathbb{X}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{\phi_p(\boldsymbol{x})} d\boldsymbol{x}, \end{aligned} \tag{12}$$

where $H(\phi_p) = 0.5 \log |\mathbf{S}| + \frac{d}{2}\ \log 2\pi + \frac{d}{2}$ is the d-dimensional entropy of normal estimate ϕ_p, where $|\mathbf{S}|$ is the determinant of a covariance matrix $\mathbf{S}$; and $J(p)$ is called negentropy, which measures the distance to normal distribution. From Eq. (11), $p(\boldsymbol{x}) = \phi_p(\boldsymbol{x})[1 + Z(\boldsymbol{x})]$, where $Z(\boldsymbol{x}) = \frac{1}{3!}\sum_{i,j,k} \kappa^{i,j,k} h_{i,j,k}(\boldsymbol{x})$. Substituting $p(\boldsymbol{x})$ in Eq. (12) leads to

$$H(p) \approx H(\phi_p) - \int_{\mathbb{X}} \phi_p(\boldsymbol{x})[Z(\boldsymbol{x}) + 0.5 Z(\boldsymbol{x})^2] d\boldsymbol{x}.$$

Using $\int_{\mathbb{X}} \phi_p(\boldsymbol{x}) Z(\boldsymbol{x}) d\boldsymbol{x} = 0$ and the orthogonal properties of Hermite polynomials, i.e., $\int_{-\infty}^{\infty} \phi_p(\boldsymbol{x}) h_n(\boldsymbol{x}) h_m(\boldsymbol{x}) d\boldsymbol{x} = n! \delta_{nm}$, where δ_{nm} is the Kronecker delta, Hulle [6] obtained an approximate expression for $H(p)$

$$\begin{aligned} H(p) \approx H(\phi_p) \quad - \quad & \frac{1}{12}\sum_{i=1}^{d} (\kappa^{i,i,i})^2 - \frac{1}{4}\sum_{i,j=1, i\neq j}^{d} (\kappa^{i,i,j})^2 \\ & - \frac{1}{72}\sum_{i,j,k=1, i<j<k}^{d} (\kappa^{i,j,k})^2. \end{aligned} \tag{13}$$

We utilize Eq. (13) for the estimation of $\hat{H}(X)$, $\hat{H}(Y)$, and $\hat{H}(X,Y)$ and substitute in Eq. (2) to get the MI estimates.

3.2.2.4 Adaptive partitioning of the XY plane (Cellucci) Cellucci et al. [8] developed a procedure for estimating MI such that the null hypothesis, i.e., H_0: X and Y are statistically independent, is rejected. They used an adaptive partitioning of the XY plane to estimate the joint probability density, i.e., $\hat{p}_{XY}$. The XY plane is nonuniformly partitioned in such a way that the Cochran criterion on $E_{XY}(i,j)$, i.e., $E_{XY}(i,j) \geq 5$ for at least 80% of all elements, is satisfied, where $E_{XY}(i,j)$ is the expected number of points in the (i,j)th element of the XY partition given the assumption of X and Y being statistically independent is valid. The whole procedure of Cellucci et al. [8] is described below.

Let x and y axes be partitioned into equal number of elements denoted by N_E which leads to

$$\hat{p}_X(i) = \hat{p}_Y(j) = \frac{n/N_E}{n},\ for\ i,j = 1,\ldots,n$$

where n is the total number of points and $\hat{p}_X(i)$ and $\hat{p}_Y(j)$ are the marginal densities at ith element of the x axis and jth element of the y axis, respectively. Under the null hypothesis that X and Y are statistically independent, the expected number of points in the (i,j)th element of the XY partition is given as

$$E_{XY}(i,j) = n\hat{p}_X(i)\hat{p}_Y(j) = \frac{n}{{N_E}^2}.$$

N_E is computed from a more conservative criterion, i.e., $E_{XY}(i,j) = n/{N_E}^2 \geq 5$ for all elements, rather than the Cochran criterion. After computing N_E, N_E partitions in the x axis and N_E partitions in the y axis are used for the estimation of joint probability density at the (i,j)th element of the XY partition, i.e., $\hat{p}_{XY}(i,j)$. The MI is estimated by substituting $\hat{p}_X$, $\hat{p}_Y$, and $\hat{p}_{XY}$ in the equation given as

$$\hat{I}(X;Y) = \sum_{i=1}^{N_E}\sum_{j=1}^{N_E} \hat{p}_{XY}(i,j)\log\frac{\hat{p}_{XY}(i,j)}{\hat{p}_X(i)\hat{p}_Y(j)}.$$

3.3 Details of the data

We analyze simple examples of linear, quadratic, and periodic functions, as well as a chaotic system, specifically the Henon map, contaminated with different levels of artificial Gaussian noise.

- *Linear:* A simple linear function with Gaussian noise can be generated as

$$X \sim N(0,1),\ Y : y_i = x_i + \varepsilon_i,$$

where $i = 1, \ldots, n$, and X is independent and identically distributed (iid). $\varepsilon \sim N(0, \sigma_\varepsilon)$ is the Gaussian noise with zero mean and standard deviation σ_ε. In this case, σ_ε gives the noise level. ε is *iid* and independent of X.

- *Quadratic:* We generate a simple quadratic, with artificial Gaussian noise, in the following manner

$$X \sim N(0,1),\ Y : y_i = {x_i}^2 + \varepsilon_i,$$

where $i = 1, \ldots, n$; X is *iid* and $\varepsilon \sim N(0, \sigma_\varepsilon)$ is the Gaussian noise with zero mean and standard deviation σ_ε. ε is *iid* and independent of X.

- *Periodic:* We consider a simple periodic function, specifically the sine function, contaminated with Gaussian noise in the following way

$$X \sim Uniform(-\pi, \pi),\ Y : y_i = \sin(x_i) + \varepsilon_i,$$

where $i = 1, \ldots, n$, and X is uniformly distributed between -π to π. $\varepsilon \sim N(0, \sigma_\varepsilon)$ is the Gaussian noise with zero mean and standard deviation σ_ε. ε is *iid* and independent of X.

- *Chaotic:* We consider the Henon map given as

$$H_X : H_{x_{i+1}} = 1 - \alpha {H_{x_i}}^2 + H_{y_i},$$
$$H_Y : H_{y_{i+1}} = \beta H_{x_i},$$

where $i = 1, \ldots, n$; $\alpha = 1.4; \beta = 0.3$; and $(H_{x_1}, H_{y_1}) = (0.0, 0.0)$. The Henon map contaminated with Gaussian noise is generated as

$$X : x_i = H_{x_i} + \varepsilon_{x_i},\ Y : y_i = H_{y_i} + \varepsilon_{y_i},$$

where $\varepsilon_x \sim N(0, \sigma_{H_X})$ and $\varepsilon_y \sim N(0, \sigma_{H_Y})$ are *iid* and independent of H_X and H_Y, respectively. σ_{H_X} and σ_{H_Y} are the standard deviations of H_X and H_Y, respectively.

3.3.1 Computations of theoretical mutual information

In this dissertation, we consider four different types of simulations, i.e., linear, quadratic, periodic and chaotic system. For linear, quadratic, and periodic cases, the exact MIs as defined by Eq. (2) can be computed as shown below.

3.3.1.1 Linear Let $X \sim N(0,1)$, $Y : y_i = x_i + \varepsilon_i$, where $i = 1, \ldots, n$; X is independent and identically distributed (iid); and $\varepsilon \sim N(0, \sigma_\varepsilon)$, where σ_ε is the noise level, is *iid* and independent of X. Let $Z = \varepsilon$, so $Y = X + Z$. Therefore, $H(Y|X)$ can be obtained as

$$H(Y|X) = H(Z) = 0.5 \log(2\pi e {\sigma_\varepsilon}^2).$$

The probability density function (*pdf*) of Z is

$$p_Z(z) = (2\pi)^{-1/2} (\sigma_\varepsilon)^{-1} \exp\left(\frac{-z^2}{2{\sigma_\varepsilon}^2}\right).$$

The *pdf* of X is given as

$$p_X(x) = (2\pi)^{-1/2} (\sigma_X)^{-1} \exp\left(\frac{-x^2}{2{\sigma_X}^2}\right),$$

where σ_X, which is called the signal level, is the standard deviation of X.

In order to compute $H(Y)$, the *pdf* of Y, i.e., $p_Y(y)$, is needed. Since $Y = X + Z$, and X and Z are independent, $p_Y(y)$ can be obtained through the convolution of the *pdfs* of X and Z given as

$$p_Y(y) = \int_{-\infty}^{\infty} p_X(x) p_Z(y - x) dx. \tag{14}$$

Solving Eq. (26), we get

$$p_Y(y) = (2\pi)^{-1/2} ({\sigma_X}^2 + {\sigma_\varepsilon}^2)^{-1/2} \exp\left(\frac{-x^2}{2({\sigma_X}^2 + {\sigma_\varepsilon}^2)}\right).$$

Therefore, $H(Y)$ can be given as

$$H(Y) = \int p_Y(y) \log p_Y(y) dy = 0.5 \log[2\pi e({\sigma_X}^2 + {\sigma_\varepsilon}^2)].$$

Substituting $H(Y|X)$ and $H(Y)$ in Eq. (2), we get

$$I(X;Y) = 0.5 \log \left(1 + \frac{{\sigma_X}^2}{{\sigma_\varepsilon}^2}\right).$$

3.3.1.2 Quadratic Let $X \sim N(0,1)$, $Y : y_i = {x_i}^2 + \varepsilon_i$, where $i = 1, \ldots, n$; X is *iid*; and $\varepsilon \sim N(0, \sigma_\varepsilon)$, where σ_ε is the noise level, is *iid* and independent of X. Let $U = X^2$ and $Z = \varepsilon$, so $Y = U + Z$. Therefore, $H(Y|X)$ can be obtained as

$$H(Y|X) = H(Z) = 0.5 \log(2\pi e {\sigma_\varepsilon}^2).$$

The *pdf* of Z is given as

$$p_Z(z) = (2\pi)^{-1/2} (\sigma_\varepsilon)^{-1} \exp\left(\frac{-z^2}{2{\sigma_\varepsilon}^2}\right).$$

The *pdf* of U is given as

$$p_U(u) = \begin{cases} (2\pi)^{-1/2} (u)^{-1/2} \exp(\frac{-u}{2}), & u > 0 \\ 0, & otherwise \end{cases}$$

In order to compute $H(Y)$, the *pdf* of Y, i.e., $p_Y(y)$, is needed. Since $Y = U + Z$, and U and Z are independent, $p_Y(y)$ can be obtained through the convolution of the *pdfs* of U and Z given as

$$p_Y(y) = \int_{-\infty}^{\infty} p_U(u) p_Z(y-u) du. \tag{15}$$

$H(Y)$ is computed as $H(Y) = \int p_Y(y) \log p_Y(y) dy$, where $p_Y(y)$ in Eq. (27) is solved using numerical integration for different values of σ_ε. We obtain $I(X;Y)$ by substituting $H(Y|X)$ and $H(Y)$ in Eq. (2).

3.3.1.3 Periodic Let $X \sim Uniform(-\pi, \pi)$, $Y : y_i = \sin(x_i) + \varepsilon_i$, where $i = 1, \ldots, n$; X is uniformly distributed between -π to π; and $\varepsilon \sim N(0, \sigma_\varepsilon)$, where σ_ε is the noise level, is *iid* and independent of X. Let $V = \sin(X)$ and $Z = \varepsilon$, so $Y = V + Z$. Therefore, $H(Y|X)$ can be obtained as

$$H(Y|X) = H(Z) = 0.5 \log(2\pi e {\sigma_\varepsilon}^2).$$

The *pdf* of Z is given as

$$p_Z(z) = (2\pi)^{-1/2} (\sigma_\varepsilon)^{-1} \exp\left(\frac{-z^2}{2{\sigma_\varepsilon}^2}\right).$$

The *pdf* of V is given as

$$p_V(v) = (\pi)^{-1}(1-v^2)^{-1/2} \; for \; 0 \leq v < 1.$$

In order to compute $H(Y)$, the *pdf* of Y, i.e., $p_Y(y)$, is needed. Since $Y = V + Z$, and V and Z are independent, $p_Y(y)$ can be obtained through the convolution of the *pdfs* of V and Z given as

$$p_Y(y) = \int_{-\infty}^{\infty} p_V(v) p_Z(y-v) dv. \tag{16}$$

$H(Y)$ is computed as $H(Y) = \int p_Y(y) \log p_Y(y) dy$, where $p_Y(y)$ in Eq. (28) is solved using numerical integration for different values of σ_ε. We obtain $I(X;Y)$ by substituting $H(Y|X)$ and $H(Y)$ in Eq. (2).

3.4 Results

We first estimate MI from KDE, KNN, Edgeworth, and Cellucci, and then substitute in Eq. (3) to get the nonlinear CC estimates. Linear CCs are obtained from LR whereas rank-based CCs are estimated from Kendall's τ. The mean of CCs and its 90% confidence bounds are evaluated using bootstrapping. The total number of bootstrap samples used for 50, 100, and 1 000 data points are 200, 100, and 10, respectively. The correlation coefficient presented here is the mean of bootstrap samples. The lower and upper bounds of 90% confidence bounds are given as 5% and 95% quantiles of bootstrap samples, respectively.

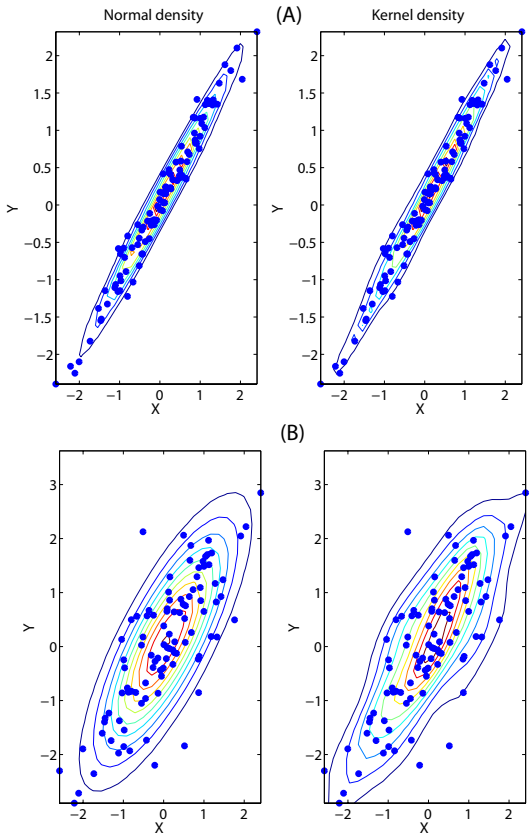

Figure 2. Linear: normal (left) and kernel (right) densities with different noise-to-signal ratios (σ_ϵ/σ_s) with 100 points. For kernel density, a Gaussian kernel with optimal smoothing parameter h_o given in Eq. (8) is used. (A) $\sigma_\epsilon/\sigma_s = 0.2$. (B) $\sigma_\epsilon/\sigma_s = 0.9$. The linear dependence structure can be seen clearly in (A) but cannot be readily identified in (B) based on eye estimation.

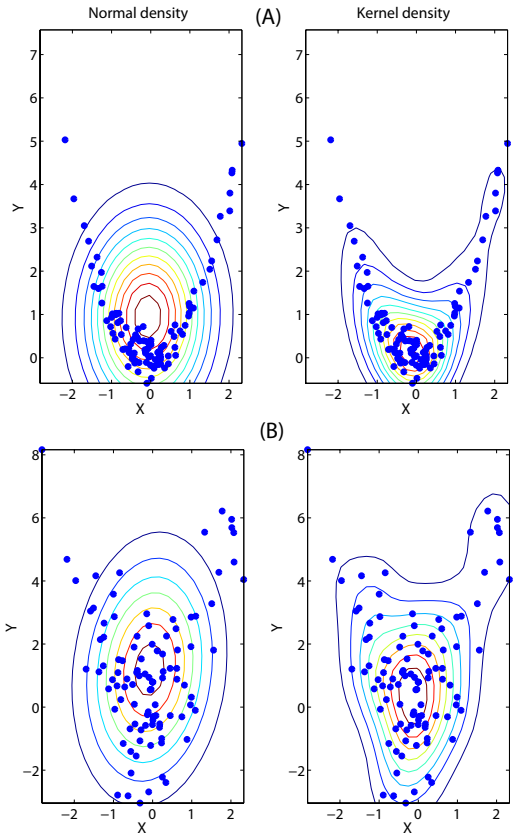

Figure 3. Quadratic: normal (left) and kernel (right) densities with different noise-to-signal ratios (σ_ϵ/σ_s) with 100 points. For kernel density, a Gaussian kernel with optimal smoothing parameter h_o given in Eq. (8) is used. (A) $\sigma_\epsilon/\sigma_s = 0.2$. (B) $\sigma_\epsilon/\sigma_s = 0.9$. At low noise, such as in (A), the nonlinear dependence can be clearly seen as shown by the kernel density. However at high noise, such as in (B), the dependence structure is not readily discernible visually from the kernel density.

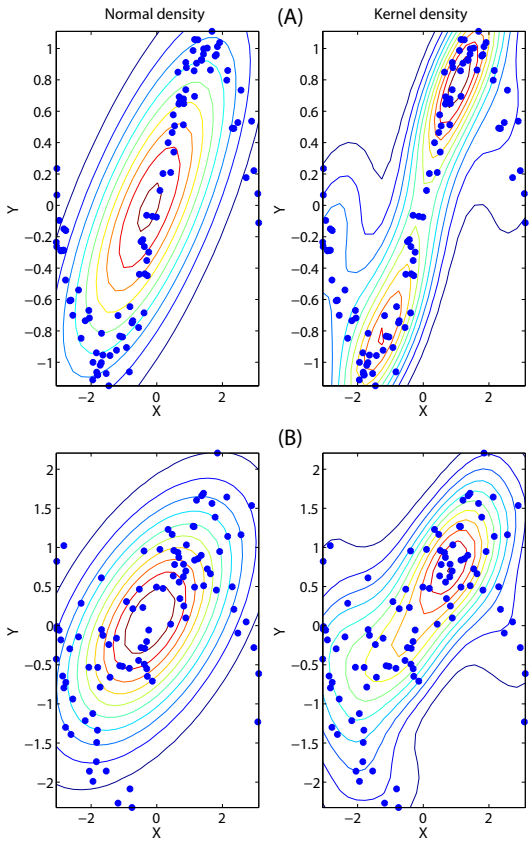

Figure 4. Periodic: normal (left) and kernel (right) densities with different noise-to-signal ratios (σ_ϵ/σ_s) with 100 points. For kernel density, a Gaussian kernel with optimal smoothing parameter h_o given in Eq. (8) is used. (A) $\sigma_\epsilon/\sigma_s = 0.2$. (B) $\sigma_\epsilon/\sigma_s = 0.9$. With increasing noise levels, the nonlinear dependence structure cannot be identified visually as shown by the kernel density plots.

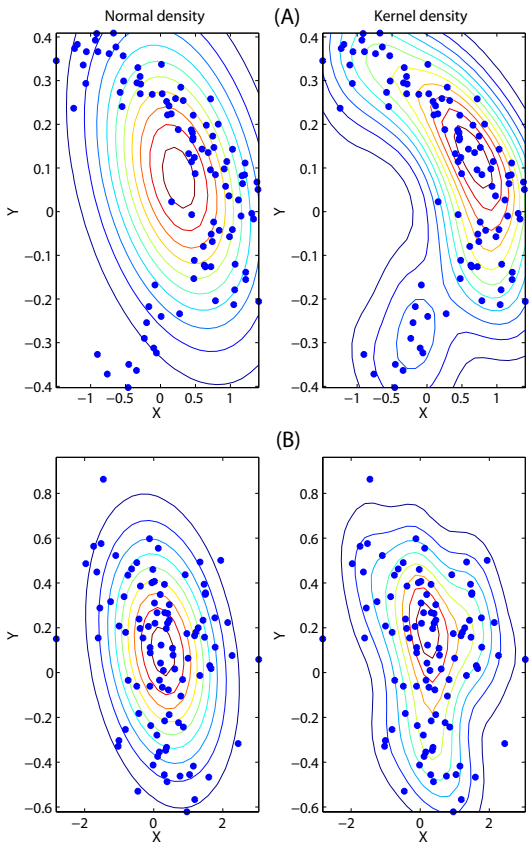

Figure 5. Chaotic: normal (left) and kernel (right) densities with different noise-to-signal ratios (σ_ϵ/σ_s) with 100 points. For kernel density, a Gaussian kernel with optimal smoothing parameter h_o given in Eq. (8) is used. (A) $\sigma_\epsilon/\sigma_s = 0.2$. (B) $\sigma_\epsilon/\sigma_s = 0.9$. Kernel density plot shows the Henon attractor in (A). However the Henon attractor cannot be readily distinguished visually in (B).

3.4.1 Performance of linear and nonlinear dependence measures

In order to compare the performance of different methods, we compare nonlinear CCs from KDE, KNN, Edgeworth, and Cellucci with linear CCs obtained from LR. If the confidence bounds of nonlinear CCs overlap with the bounds of linear CCs, it means here that nonlinear correlations are not different from linear correlations at 90% confidence level. Nonlinear CCs obtained from the MI estimation methods are compared with theoretical CCs derived from the theoretical MI values which can be computed analytically for three out of four test cases considered here, namely linear, quadratic, and periodic. The performance of the MI estimation methods is also compared with a rank-based correlation measure, specifically the Kendall's τ. Plots of normal and kernel density estimates for linear, quadratic, periodic, and chaotic are shown in Figs. 13-16.

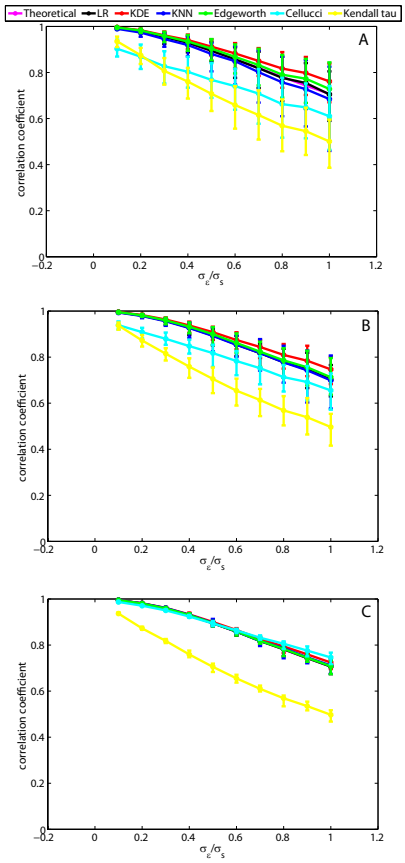

Figure 6. Linear: Comparisons between linear CCs from LR and nonlinear CCs from KDE, KNN, Edgeworth, Cellucci, and Kendall's τ, at different noise-to-signal ratios (σ_ϵ/σ_s) for (A) 50 points, (B) 100 points, and (C) 1 000 points.

Table 1. Linear: Description of results where each entry consists of three columns given as (1) Column 1: 0, -, or +, where '0', '-' and '+' mean nonlinear CCs are zero, negatively and positively biased with respect to theoretical CCs, respectively, (2) Column 2: Y or N, where 'Y' and 'N' mean 90% confidence bounds of nonlinear CCs overlap and do not overlap with theoretical CCs, respectively, and (3) Column 3: Y or N, where 'Y' and 'N' mean 90% confidence bounds of nonlinear CCs overlap and do not overlap with linear CCs, respectively. *Bold* and *slanted* entries indicate the best and the second best methods for each case specified in the top headings of the table, respectively.

	Very short data		Short data	
	low noise	high noise	low noise	high noise
KDE	*+ Y Y*	**+ Y Y**	0 Y Y	+ Y Y
KNN	**0 Y Y**	- Y Y	**0 Y Y**	**0 Y Y**
Edgeworth	+ Y Y	*+ Y Y*	*0 Y Y*	*0 Y Y*
Cellucci	- N N	- Y Y	- Y Y	+ Y Y
Kendall's τ	- N N	- N Y	- N N	- N N

3.4.1.1 Linear Linear and nonlinear CCs with 90% confidence bounds are shown in Fig. 17. The theoretical CC, which is computed analytically, is expected to be identical to the linear CC. As noise levels increase, linear and nonlinear CCs decrease and their corresponding variances increase for both *very short* and *short* data. The complete description of results obtained from KDE, KNN, Edgeworth, Cellucci, and Kendall's τ for *very short* and *short* data at low and high noise is given in Table 1. For *very short* data, KNN seems to be a better choice at low noise because it has no bias, overlaps with theoretical CCs and has narrow confidence bounds (Figs. 17A and 17B). At high noise, KDE is positively biased but it appears to be a better choice given that the others have wider confidence bounds. Thus, for *very short* data, KNN may be utilized at low noise but at high noise, KDE seems to be the best choice. For *short* data, Kendall's τ is the worst whereas Edgeworth is better than KDE because it overlaps exactly with theoretical CCs (Fig. 17C). LR and KNN stand out among the rest since they have very small bias, overlap exactly with theoretical CCs and have narrow bounds at all noise levels. Thus, for *short* data, either KNN or LR may be utilized at all noise levels.

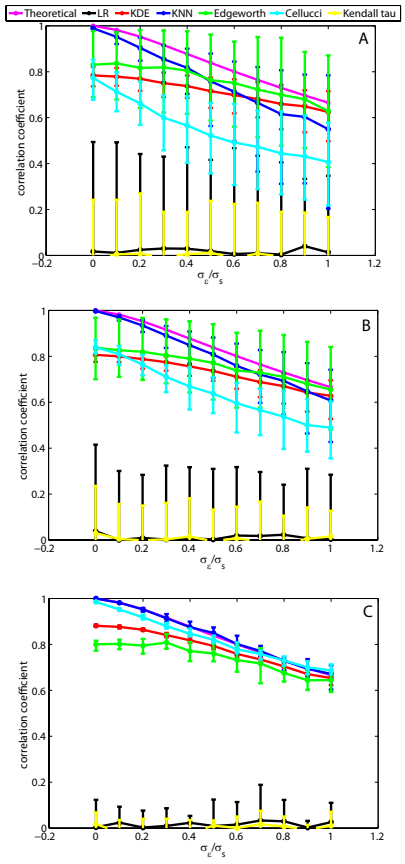

Figure 7. Quadratic: Comparisons between linear CCs from LR and nonlinear CCs from KDE, KNN, Edgeworth, Cellucci, and Kendall's τ, at different noise-to-signal ratios (σ_ϵ/σ_s) for (A) 50 points, (B) 100 points, and (C) 1 000 points.

Table 2. Quadratic: Description of results where each entry consists of three columns given as (1) Column 1: 0, -, or +, where '0', '-' and '+' mean nonlinear CCs are zero, negatively and positively biased with respect to theoretical CCs, respectively, (2) Column 2: Y or N, where 'Y' and 'N' mean 90% confidence bounds of nonlinear CCs overlap and do not overlap with theoretical CCs, respectively, and (3) Column 3: Y or N, where 'Y' and 'N' mean 90% confidence bounds of nonlinear CCs overlap and do not overlap with linear CCs, respectively. *Bold* and *slanted* entries indicate the best and the second best methods for each case specified in the top headings of the table, respectively.

	Very short data		Short data	
	low noise	high noise	low noise	high noise
KDE	- N N	**- Y N**	- N N	- Y N
KNN	**- Y N**	- Y Y	**0 Y N**	**0 Y N**
Edgeworth	- *Y N*	- *Y N*	- N N	- Y N
Cellucci	- N N	- Y Y	- *N N*	- *Y N*
Kendall's τ	- N Y	- N Y	- N Y	- N Y

3.4.1.2 Quadratic LR and Kendall's τ fail to capture the nonlinear dependence as shown by near zero CC in Fig. 21. The variance increases for KDE, KNN, Edgeworth, and Cellucci as the noise level increases at all noise levels. Table 2 gives the complete description of results obtained from LR, KDE, KNN, Edgeworth, Cellucci, and Kendall's τ for *very short* and *short* data at low and high noise. For *very short* data, as the noise level increases, the bias increases for KNN and Cellucci and decreases for KDE and Edgeworth (Figs. 21A and 21B). At low noise, only KNN and Edgeworth overlap with theoretical CCs but KNN is more closer to theoretical than Edgeworth. At high noise, the performance of KDE is the best because it is closer to theoretical CCs, does not intersect with linear CCs, and has narrow confidence bounds as compared to that from KNN, Edgeworth, and Cellucci. Thus, for *very short* data, KNN and KDE may be utilized at low and high noise, respectively. For *short* data, KNN is the best because it overlaps exactly with theoretical CCs and has narrow confidence bounds (Fig. 21C). Cellucci is more closer to theoretical CCs than KDE and Edgeworth. Thus, KNN seems to be the best choice for *short* data. KDE may be further improved at low noise by choosing a smaller value of the smoothing parameter.

Table 3. Periodic: Description of results where each entry consists of three columns given as (1) Column 1: 0, -, or +, where '0', '-' and '+' mean nonlinear CCs are zero, negatively and positively biased with respect to theoretical CCs, respectively, (2) Column 2: Y or N, where 'Y' and 'N' mean 90% confidence bounds of nonlinear CCs overlap and do not overlap with theoretical CCs, respectively, and (3) Column 3: Y or N, where 'Y' and 'N' mean 90% confidence bounds of nonlinear CCs overlap and do not overlap with linear CCs, respectively. ***Bold*** and *slanted* entries indicate the best and the second best methods for each case specified in the top headings of the table, respectively.

	Very short data		Short data	
	low noise	high noise	low noise	high noise
KDE	- *N Y*	**- Y Y**	- N N	- *N N*
KNN	**- Y N**	- *Y Y*	**0 Y N**	**0 Y N**
Edgeworth	- N Y	- N Y	- N Y	- N Y
Cellucci	- N Y	- Y Y	- *N N*	+ N N
Kendall's τ	- N N	- N Y	- N N	- N N

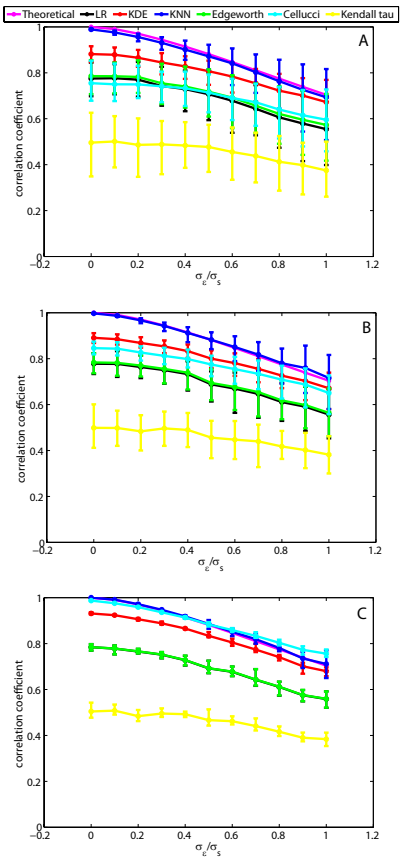

Figure 8. Periodic: Comparisons between linear CCs from LR and nonlinear CCs from KDE, KNN, Edgeworth, Cellucci, and Kendall's τ, at different noise-to-signal ratios (σ_ϵ/σ_s) for (A) 50 points, (B) 100 points, and (C) 1 000 points. In (C), LR overlaps exactly with Edgeworth.

3.4.1.3 Periodic Correlation coefficients and their 90% confidence bounds obtained from LR, KDE, KNN, Edgeworth, Cellucci, and Kendall's τ are shown in Fig. 22. KNN overlaps with theoretical CCs for both *very short* and *short* data at all noise levels except for the fact that at high noise it produces wide confidence bounds. The performance of Kendall's τ is the worst at all noise levels. Edgeworth appears to capture only the linear correlation and produces wide confidence bounds. In this case the density of Y is bimodal, which causes Edgeworth estimates to be incorrect. The results obtained from LR, KDE, KNN, Edgeworth, Cellucci, and Kendall's τ are described in Table 3 for *very short* and *short* data at low and high noise. For *very short* data, the variances from all the methods are small at low noise but increase as the noise level increases (Figs. 22A and 22B). KNN and KDE have the lowest variances at low and high noise, respectively. KNN overlaps exactly with theoretical CCs and has narrow and wide confidence bounds at low and high noise, respectively. Thus, KNN is a better choice at low noise. At high noise, KDE and KNN overlap with theoretical CCs as well as with linear CCs but KDE has the smallest confidence bounds. Thus, for *very short* data, KNN and KDE may be utilized at low and high noise, respectively. For *short* data, there is not much difference in the variances from all methods with Cellucci having the lowest variance (Fig. 22C). The performances of KNN and Cellucci are better than the rest. Cellucci overlaps with theoretical CCs for only few noise levels whereas KNN overlaps exactly with theoretical CCs and has narrow bounds. Thus, KNN has an edge over all other methods considered here for *short* data across all noise levels.

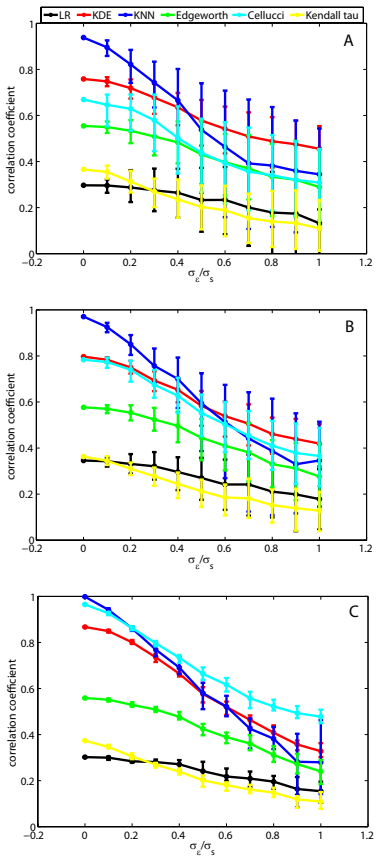

Figure 9. Chaotic: Comparisons between linear CCs from LR and nonlinear CCs from KDE, KNN, Edgeworth, Cellucci, and Kendall's τ, at different noise-to-signal ratios (σ_ϵ/σ_s) for (A) 50 points, (B) 100 points, and (C) 1 000 points.

3.4.1.4 Chaotic For *very short* and *short* data, linear CCs between X and Y components of the Henon map are negative for all noise levels. Since nonlinear CCs from the MI estimation methods do not have directionality, the absolute values of linear CC are considered here. Note that theoretical CCs for the Henon map could not be computed analytically and were not found in the literature. However, given the dynamical relation between X and Y, nonlinear CCs are expected to be greater than linear CCs at all noise levels. Nonlinear and linear CCs decay as noise level increases.

For *very short* data, KNN estimates higher CCs than all other methods when σ_ϵ/σ_s is less than around 0.5 after which KDE yields higher values compared to all other methods (Figs. 23A and 23B). The performance of Kendall's τ is the worst since it captures less dependence than the linear correlation for the majority of noise levels. At low noise, both Edgeworth and Cellucci are ruled out because they are lower than KNN and KDE and have wide confidence bounds. Thus, KNN seems to be a better choice at low noise since KDE is negatively biased. As the noise level increases, the confidence bounds from all methods increase. At high noise, the confidence bounds from KNN, Edgeworth, and Cellucci overlap with linear CCs. KDE seems to have an edge over the other methods since it has narrow confidence bounds and does not overlap with linear CCs. Thus, KNN and KDE may be utilized for *very short* data at low and high noise, respectively. For *short* data, Cellucci differs completely from the other estimators at high noise (Fig. 23C). KNN is a better choice at low noise because it appears to be the most consistent. At high noise, KNN and Edgeworth are ruled out because they overlap with linear CCs due to their wide confidence bounds. KDE overlaps with KNN but it stands out due to its ability to capture more correlation than purely linear correlation. Thus, for *short* series, KNN and KDE may be utilized at low and high noise, respectively.

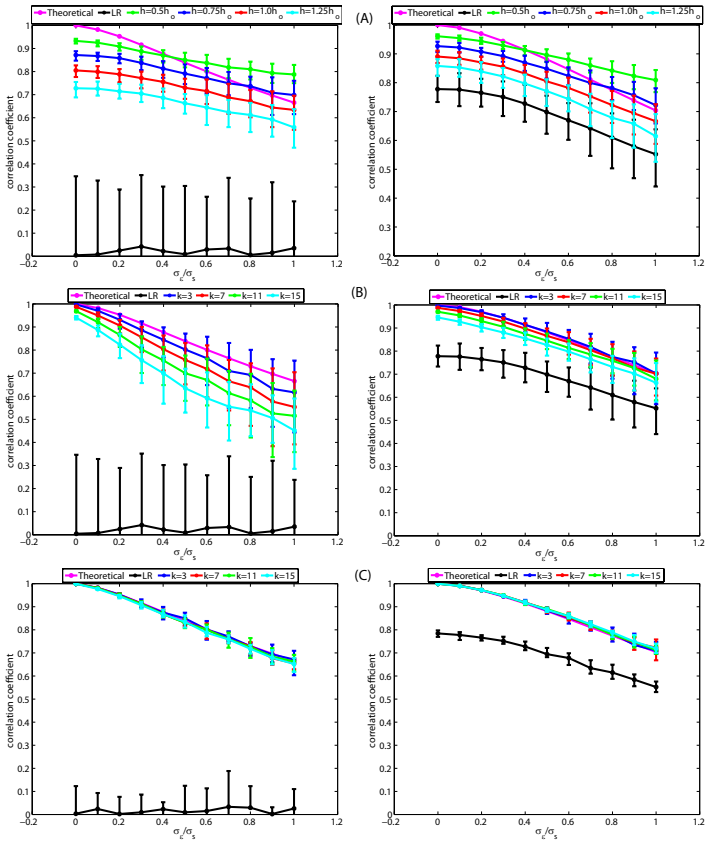

Figure 10. Performance of KDE and KNN with different values of smoothing parameter (h) and number of nearest neighbors (k), respectively. The results from quadratic and periodic are presented in the left and right, respectively. (A) KDE with 100 points. (B) KNN with 100 points. (C) KNN with 1 000 points. In (A), h_o is the optimal smoothing parameter for a Gaussian kernel given in Eq. (8).

3.4.2 Performance of KDE and KNN with different parameter values

In the case of KDE, the amount of smoothing defined by smoothing parameter, h in Eq. (5), is very important for the density estimation, which, in turn, influences the MI estimates. The selection of appropriate smoothing parameter needs to be guided by the end-use of the density estimates. Here we use the optimal smoothing parameter for a Gaussian kernel (h_o) with KDE given in Eq. (8). We investigate the effects of h on nonlinear CC estimates from KDE by selecting different values of h around h_o. For KNN, the number of nearest neighbors (k) governs the overall amount of smoothing in the densities which are subsequently used in entropy estimation given in Eq. (10). Small values of k lead to small bias and large variance whereas large k results in large bias and small variance. Thus, the bias-variance tradeoff, which is a common issue encountered in statistical estimation procedures, is also important here. Kraskov et al. [5] warned against using large k since

the decrease in variance is outweighed by the increase in bias. They proposed k ranging from 2 to 4. Here we use k as three for KNN. We evaluate the effects of k on nonlinear CC estimates from KNN by selecting different k values. The results presented here are obtained for two cases, specifically, quadratic and periodic.

For *very short* data, the bias and variance from KDE increase with the increase of h at low noise and all noise levels, respectively (Fig. 10A). At low noise, KDE does not overlap with theoretical CCs. However, KDE with $h = 0.75h_o$ and $h = h_o$ performs better at high noise since their 90% confidence bounds overlap with theoretical CCs. The bias and variance from KNN increase as the number of nearest neighbors increase across all noise levels (Fig. 10B). At low noise, the performance of KNN with $k = 3$ is the best of all the cases considered here since it has small bias and its confidence bounds overlap with theoretical CCs. At high noise, KNN has large bias and variance for all k. If KNN needs to be used at high noise, $k = 3$ appears to be a better choice since it is closer to theoretical CCs as compared to the others and the variances from all k are comparable. Thus, for *very short* data, KDE with $h = 0.75h_o$ or $h = h_o$ may be utilized at high noise whereas KNN with $k = 3$ seems to be a better choice at low noise.

For *short* data, KNN with $k = 3$ performs better at low noise since it has small bias and variance (Fig. 10C). As k increases, the bias increases and the variance decreases at high noise. KNN with all k considered here performs better at high noise but the selection of appropriate k needs to be guided by the acceptable levels of bias and variance. Thus, for *short* data, KNN with $k = 3$ is the best since it overlaps exactly with theoretical CCs and its variance does not differ significantly from the others.

3.5 Conclusion and discussion

Our results indicate that two MI-estimation methods, specifically KDE and KNN, outperform the other methods and estimation procedures in terms of their ability to capture the dependence structure including nonlinear dependence where present. We find that KNN is the best estimator for *very short* data with relatively low noise while KDE works better for *very short* data when the noise levels are higher. A visual examination of the density plots may help in explaining the relative performance of KDE and KNN (Figs. 13-16 in *Appendix B*). For *short* data, KNN is the best choice for capturing the nonlinear dependence across all noise levels except when the data are generated from chaotic dynamics, where KDE is a better choice at higher noise levels. We surmise that the relative performance of KDE and KNN with respect to various noise levels is a consequence of the bias-variance tradeoff. Previous literature suggests that KDE estimates can often be highly biased if the particular KDE recipe used here is followed [85], while KNN estimates can have significant variance when the number of nearest neighbors (k) is set to low values, e.g., $k = 3$ as used in this

dissertation. The bias in the KDE estimates dominates the variance of the estimates for low noise-to-signal ratios. The KNN performs relatively better for low noise levels since its bias and variance are lower than that from KDE. However, the converse is true for high noise-to-signal ratios, and hence the KDE performs relatively better. For high noise, the variance dominates because of the noise in the data but the variance associated with $k = 3$ for KNN increases dramatically. One way to address the large variance from KNN is to use a much larger value of k but it would also increase the bias.

In general, the above discussions and pointers appear to suggest that the results for nonlinear dependence obtained from KDE and KNN may in fact reflect the lower bounds of what may be potentially achievable through improvements or intelligent combinations of KNN and KDE. Specifically, both the KDE and KNN estimates can be potentially improved by utilizing a plug-in method for kernel, smoothing parameter (h), or k selection. Such plug-in procedures would cause additional estimation variance but may reduce the overall MSE of estimation. However, the development or utilization of procedures for the selection of optimal kernels, smoothing parameters, or nearest neighbors, may be rather involved and hence is left as an area of future research.

We have presented preliminary justifications for the relative performance of the MI estimation methods based on considerations like the bias-variance tradeoff and the nature of the approximations underlying the estimation procedures. Our evaluation suggests that the development of guidance for the use of the most suitable estimation procedure may be possible and would depend on known data or domain characteristics and exploratory data analysis. If such guidance can indeed be provided, this could conceivably lead to the development of automated or semi-automated procedures for the choice of the most appropriate estimation procedure and the corresponding parameters. However, significant future research on multiple test cases comprising simulated and real data may be necessary before such procedures can be deployed in real world settings.

Acknowledgements

This research was partially funded by the Laboratory Directed Research and Development Program of the Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. The authors are thankful to Dr. Alexander Kraskov for providing the KNN-based mutual information code. The authors would like to thank two anonymous reviewers for their helpful suggestions which significantly improved the quality of the paper. Shiraj Khan was partially funded by the Department of Civil and Environmental Engineering (CEE) at the University of South Florida

(USF) in Tampa, FL. Auroop R. Ganguly acknowledges partial funding from the $SensorNet^{®}$ research program at ORNL. The authors would like to acknowledge Drs. Gabriel Kuhn, Olufemi Omitaomu and Ranga Raju Vatsavai, of ORNL for their helpful comments. The United States Government and the publisher, by accepting the article for publication, acknowledge that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for its purposes.

Chapter 4

Nonlinear Statistics Reveals Stronger Ties Between ENSO and the Tropical Hydrological Cycle

Cross-spectrum analysis based on linear correlations in the time domain suggested a coupling between large river flows and the El Niño-Southern Oscillation (ENSO) cycle. A nonlinear measure based on mutual information (MI) reveals extrabasinal connections between ENSO and river flows in the tropics and subtropics, that are 20-70% higher than those suggested so far by linear correlations. The enhanced dependence observed for the Nile, Amazon, Congo, Paraná, and Ganges rivers, which affect large, densely populated regions of the world, has significant impacts on inter-annual river flow predictabilities and, hence, on water resources and agricultural planning.

4.1 Introduction

ENSO events impact regional precipitation in the tropics and subtropics, ultimately causing inter-annual variability in river flows. The ocean-atmosphere-land interactions are complex and far from being completely understood and accurately modeled. A slight disturbance in these interactions would usually result in sometimes surprising distant correlations and climate patterns. Analyses of the rainfall anomalies during the warm (El Niño) and cold (La Niña) episodes of ENSO suggest the existence of nonlinear sea surface temperature (SST)-rainfall relationships in the tropics and a strong influence of SST forcing on equatorial rainfall in the geographic vicinity of that forcing [11]. To properly explain and ultimately predict this variability, it is important to disentangle, as far as possible, long range climatic phenomena from recent effects such as those possibly produced by deforestation and global warming.

While the relationships among many climate and hydrological variables are decidedly nonlinear [12], *linear* dependence measures are still being used as a matter of course to relate ENSO and inter-annual variability in river flows. These measures have ranged from linear correlation coefficients (CC) in the time domain [13–16] to the cross-spectrum analysis [17, 18]. One of the reasons for using linear measures is that the inherent noise and periodicity in the observations together with short length of the available sample sizes make it difficult to use nonlinear approaches in climate and hydrology [19–21].

The goal of this dissertation is to investigate the nonlinear dependence between ENSO and the annual flow of some of the largest tropical and subtropical rivers, specifically the Nile, Amazon, Congo, Paraná and Ganges, through a measure based on the mutual information (MI). The results reveal a stronger extrabasinal connection between ENSO and river flows than the one suggested by linear analysis using linear regression (LR). This has significant impacts on scientific understanding and predictability as well as management of water and agricultural resources in vast, densely populated regions of the globe.

4.2 Data and methodology

4.2.1 ENSO and river flow data

ENSO events are associated with SST anomalies over the eastern and central equitorial Pacific Ocean. In this dissertation, the ENSO index is defined in terms of the monthly SST variations from the long-term mean, averaged over the regions $2^o - 6^o$N, $90^o - 170^o$W; 2^oN-6^oS, $90^o - 180^0$W; and $6^o - 10^o$S, $110^o - 150^o$W of the Pacific Ocean. This dataset was published as a homogenized monthly series of the mean SST anomaly for the period 1872-1989 [98]. After 1989, the NINO 3.4 is used as the ENSO index because its geographical regions (5^oN-5^oS; $120^o - 170^o$W) are close to regions corresponding to the Wright SST.

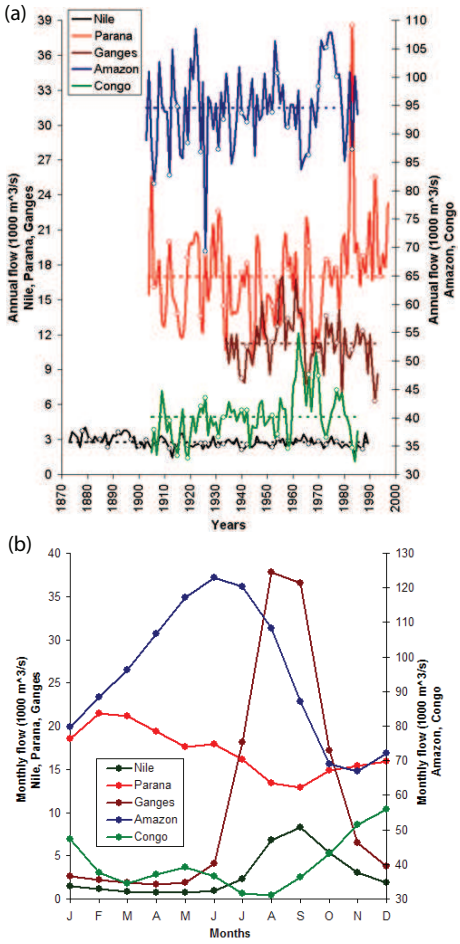

Figure 11. Annual flow (a) and average monthly flow (b) of the Nile River from 1873-1989, Amazon River from 1903-1985, Congo River from 1905-1985, Paraná River from 1904-1997, and Ganges River from 1934-1993. The following years are happened to be the warm episodes of ENSO: 1877, 1880, 1884, 1887, 1891, 1896, 1899, 1902, 1905, 1911, 1914, 1918, 1923, 1925, 1930, 1932, 1939, 1941, 1951, 1953, 1957, 1965, 1969, 1972, 1976, 1982, 1986, 1991, 1993, and 1997. The average annual flow and El Niño years are shown as dotted lines and solid dots, respectively, as shown in (a).

The monthly discharge data of the Nile River was measured at Aswan (lat. 24.1^oN, long. 33^oE) from 1873 to 1989. This integrated runoff comprises contributions from three major tributaries, i.e. the White Nile, the Blue Nile, and the Atbara, and represents the majority of the Nile basin. The seasonal streamflow cycle of the Nile indicates that the minimum and maximum discharges are observed in April and September, respectively (Figure 11b).

The discharge data of the Amazon River was collected monthly from the Rio Negro stage at Manaus (lat. 3^oS, long. 60^oW) over the period from 1903 to 1985. The integrated runoff at the Manaus gauge covers more than 3M km^2 of the Andean and western Amazon watershed [17]. The seasonal streamflow cycle of the Amazon indicates that the minimum and maximum discharges are observed in November and June, respectively (Figure 11b).

The Congo River discharge data was collected monthly from the river stage at Kinshasa, Zaire (lat. 4.3^oS, long. 15.3^oE) from 1905 to 1985. As the Congo River basin, covering approximately 3.8M km^2, is located around the equator, it experiences a marked semi-annual rainfall cycle which is associated with the north/south movement of the inter tropical convergence zone (ITCZ) across tropical Africa [99]. This is evident from the seasonal cycle of the Congo river indicating two peaks in May and December (maximum) and the lowest flow in August (Figure 11b).

The Paraná River discharge data for the period 1904-1997 was collected monthly at Corrientes (lat. 27^oS, long. 59^oW) located downstream of the confluence of the Paraguay and the Paraná rivers. The seasonal cycle of the Paraná exhibits a single peak in February with a long recession and low discharge in September (Figure 11b).

The monthly Ganges River discharge data was recorded over the period from 1934 to 1993 at the Hardinge Bridge in Bangladesh by the Bangladesh Water Development Board. It experiences the flood season from July to October, during which the average annual flow is 82% [15]. The peak flow and low discharge of the Ganges are observed in August and April, respectively, as exhibited by the seasonal cycle (Figure 11b).

Table 4. Runoff data statistics (1000 m^3/s).

Parameter	Nile	Amazon	Congo	Paraná	Ganges
Mean	2.79	113.50	481.84	204.49	134.22
Std. dev.	0.46	85.24	50.90	49.64	27.03
Max. (year)	4.06 (1879)	1301 (1922)	659.57 (1962)	462.73 (1983)	202.84 (1956)
Min. (year)	1.46 (1913)	832 (1926)	386.81 (1984)	112.95 (1944)	75.46 (1992)

During the year following the warm episodes of ENSO, the annual discharges of the Nile, Amazon, Congo and Ganges Rivers fall below their average annual discharge whereas the annual Paraná discharge is higher than the average annual discharge (Figure 11a). The runoff statistics give an idea about the discharge characteristics of the rivers (Table 4).

4.2.2 Mutual information (MI)

MI is a measure of statistical dependence among random variables which captures the full dependence structure, both linear and nonlinear. The concept of MI was originally developed in communication theory and has been applied to multiple domains over the last few decades [5,9]. Considering two random variables X and Y, the MI, denoted by $I(X;Y)$, is defined as

$$I(X;Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y), \tag{17}$$

where $H(X)$ or $H(Y)$ is the marginal information entropy which measures the information content in a signal and $H(X,Y)$ is the joint information entropy which measures the information content in a joint system of X and Y. The MI between two random variables X and Y can also be defined as

$$I(X;Y) = \int_Y \int_X p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x) p_Y(y)} \, dx \, dy, \tag{18}$$

where $p_{XY}(x,y)$ is the joint probability density function (*pdf*) between X and Y, and $p_X(x)$ and $p_Y(y)$ are the marginal *pdfs*. The MI values range from 0 (independent) to ∞ (completely dependent). For a bivariate normal set (X,Y), the MI and the linear CC, denoted by ρ, are related as $I(X;Y) = -0.5 \log[1 - \rho(X,Y)^2]$ [93]. For comparing linear and nonlinear dependence measures, the MI based nonlinear CC, i.e., λ, ranging from 0 to 1 is defined from the above relationship as

$$\hat{\lambda}(X,Y) = \sqrt{1 - \exp[-2\hat{I}(X;Y)]}, \tag{19}$$

where $\hat{\lambda}(X,Y)$ and $\hat{I}(X;Y)$ are the estimated nonlinear CC and MI between X and Y, respectively [93,94]. In addition, just as the mean squared errors (MSE) can be derived from LR, a lower bound of MI-based MSE, which is a measure of the predictability of Y based on the information content in X, can be estimated as

$$\widehat{MSE}(Y) \geq \frac{1}{2\pi e} \exp[2(\hat{H}(Y) - \hat{I}(X;Y))], \tag{20}$$

where $\hat{H}(Y)$ is the estimated entropy of Y and $\hat{I}(X;Y)$ is the estimated MI between X and Y [87]. ANOVA-like interpretations have also been suggested for MI-based dependence [87]. Cellucci et al. [8] compared MI-based dependence with traditional measures of dependence, such as Pearson linear correlation coefficient, Spearman rank order correlation, and Kendall's tau.

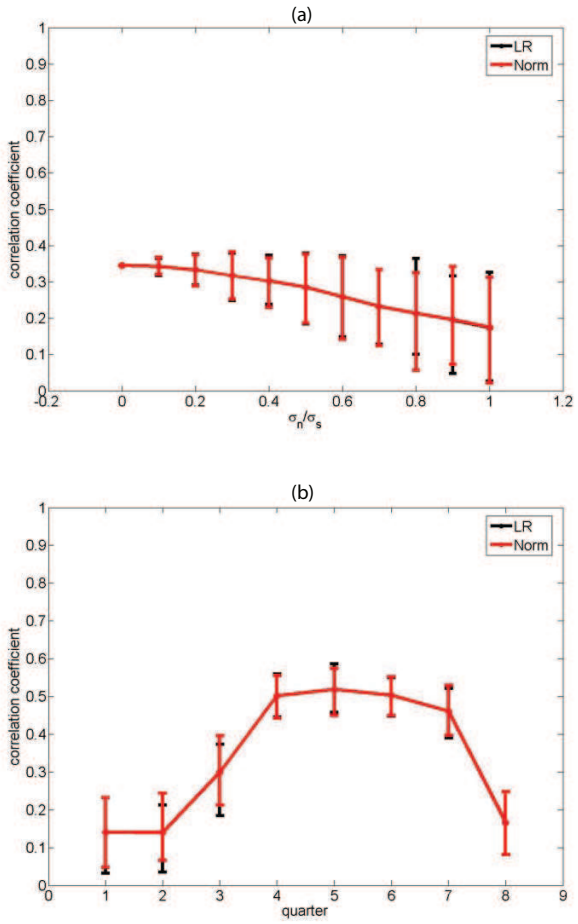

Figure 12. Comparison of linear (LR) and MI-based dependence obtained after fitting bivariate normal distribution (Norm) to each pair. The dependence is estimated with respect to different noise to signal ratios and quarters in the simulated and real data, respectively. (a) Case 4 (Chaotic): Henon map with 100 points. (b) Relationship between ENSO and Nile River flow. In (a), the mean dependence from both cases are same whereas there is a very slight difference in variances for few noise to signal ratios. In (b), both cases capture the same mean dependece whereas variances differ very slightly for few quarters.

The measures for linear (ρ) and nonlinear (λ) correlation quantify the strength of dependence among multiple variables (viz., ENSO and streamflow in this dissertation). While the former quantifies the dependence purely in terms of the linear information content and the latter quantifies the complete (linear and nonlinear) information content, the two measures can be related in principle since they both capture the information contained in one variable about the other. The relationship between the two measures (ρ and λ) has been explained in detail by Brillinger [87]. In more rigorous terms, the two measures can be compared quantitatively since they directly relate to the expected MSEs from predictions (see equation (34) for MSE from the MI-based dependence). The confidence bounds reflect the degree of belief in the two measures and hence can be compared as well. The definition of the nonlinear measure (λ) used here has been utilized by previous researchers [87, 93, 94] precisely because λ collapses to the linear measure (ρ) for the bivariate normal distribution (see equation (19)). We compare ρ from LR and λ obtained from first estimating the MI after fitting bivariate normal distribution to the data and then using equation (19). We test one simulation, i.e., chaotic (described in *section 4.1*), and one real data, i.e., dependence between ENSO and Nile River flow, and observe that λ obtained after fitting bivariate normal distribution is exactly similar to ρ for both cases (Figure 12). Finally, we would like to emphasize that the statements that compare linear and nonlinear correlation measures, while statistically valid, need to be evaluated with care owing to issues pertaining to statistical estimation like bias-variance tradeoffs.

4.3 MI estimation methods

The estimation of the MI requires the estimation of the joint and marginal *pdfs*, which, in turn, are frequently obtained from histogram and kernel density based estimators. Estimates of MI are consistent and asymptotically converge to the *true* or theoretical value when the data sets are relatively large and error-free. Since observations of river flows and the ENSO index are short and usually affected by various errors, it is important to assess various MI estimation methods for short and noisy data. Recently developed methodologies have been explored for estimating the MI, such as kernel density estimators (KDE) [2], k-nearest neighbors (KNN) [5], and Edgeworth approximation of differential entropy (Edgeworth) [6].

4.3.1 Kernel density estimator (KDE)

For any bivariate data set (X, Y) of size N, $\hat{I}(X;Y)$ is estimated as

$$\hat{I}(X;Y) = \frac{1}{N}\sum_{i=1}^{N}\log\frac{\hat{p}_{XY}(x_i, y_i)}{\hat{p}_X(x_i)\hat{p}_Y(y_i)}, \tag{21}$$

where $\hat{p}_{XY}(x_i, y_i)$ is the estimated joint *pdf*, and $\hat{p}_X(x_i)$ and $\hat{p}_Y(y_i)$ are the estimated marginal *pdfs* at (x_i, y_i).

The multivariate kernel density estimator using a normal kernel is defined as

$$\hat{p}_X(\boldsymbol{x}) = \frac{1}{Nh^d}\sum_{i=1}^{N}\frac{1}{\sqrt{(2\pi)^d|\mathbf{S}|}}\exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{x}_i)^T\mathbf{S}^{-1}(\boldsymbol{x}-\boldsymbol{x}_i)}{2h^2}\right), \tag{22}$$

where N is the number of data points; $\boldsymbol{x}$ and $\boldsymbol{x}_i$ are the d-dimensional vectors; $\mathbf{S}$ is the covariance matrix on the $\boldsymbol{x}_i$; $|\mathbf{S}|$ is the determinant of $\mathbf{S}$; and h is the kernel bandwidth also called the smoothing parameter [2]. In this dissertation, the smoothing parameter is chosen as the optimal Gaussian bandwidth for a normal kernel given as $h = [4/(d+2)]^{1/(d+4)}\ N^{-1/(d+4)}$. The MI estimates are obtained by first estimating $\hat{p}_X$, $\hat{p}_Y$, and $\hat{p}_{XY}$ from equation (22) and then plugging them in equation (21).

4.3.2 k-nearest neighbors (KNN)

The MI between X and Y is estimated as

$$\hat{I}(X;Y) = \psi(k) - \frac{1}{k} - \frac{1}{N}\sum_{i=1}^{N}[\psi(n_x(i)) + \psi(n_y(i))] + \psi(N), \tag{23}$$

where N and k are the number of data points and nearest neighbors, respectivley; if $\epsilon(i)/2$ is the distance between (x_i, y_i) and its kth neighbor, denoted by (kx_i, ky_i), and if $\epsilon_x(i)/2$ and $\epsilon_y(i)/2$ are given as $||x_i - kx_i||$ and $||y_i - ky_i||$, respectively, then $n_x(i)$ is the number of points x_j such that $||x_i - x_j|| \leq \epsilon_x(i)/2$; $n_y(i)$ can be calculated similarly; $\psi(x)$ is the digamma function, $\psi(x) = \Gamma(x)^{-1}d\Gamma(x)/dx$, which satisfies the relation $\psi(x+1) = \psi(x) + 1/x$, with $\psi(1) = -C$, where $C = 0.5772156649$ is the Euler-Mascheroni constant [5]. This dissertation chooses k as 3 since Kraskov et al. [5] suggested $k > 1$ in order to reduce statistical errors and also indicated to avoid large values of k which lead to the increase of systematic errors.

4.3.3 Edgeworth approximation of differential entropy (Edgeworth)

Using Edgeworth expansion of the density $p(\boldsymbol{x}), \boldsymbol{x} = [x_1, .., x_d]$, the differential entropy is defined as

$$\begin{aligned} H(p) &= H(\phi_p) - J(p) \\ &= H(\phi_p) - \frac{1}{12}\sum_{i=1}^{d}(\kappa^{i,i,i})^2 - \frac{1}{4}\sum_{i,j=1,i\neq j}^{d}(\kappa^{i,i,j})^2 \\ &\quad -\frac{1}{72}\sum_{i,j,k=1,i<j<k}^{d}(\kappa^{i,j,k})^2, \end{aligned} \tag{24}$$

where d is the dimension of $\boldsymbol{x}$; $H(\phi_p) = 0.5\log|\mathbf{S}| + \frac{d}{2}\log 2\pi + \frac{d}{2}$, where $\mathbf{S}$ is the covariance matrix, is the d-dimensional entropy of the best normal estimate, i.e., ϕ_p, with the same mean and covariance matrix as p; and κ is a standardized cumulant [6]. In equation (24), $J(p)$ is called negentropy, which measures the distance to normal distribution. The MI is estimated by first estimating $\hat{H}(X)$, $\hat{H}(Y)$, and $\hat{H}(XY)$ from equation (24) and then plugging them in equation (17).

4.4 Analysis of simulations

We evaluate and compare MI estimation methods, i.e., KDE, KNN, and Edgeworth, using some simulations to find the best method for the real data analysis [10]. Nonlinear CCs obtained from these methods are compared with linear and theoretical CCs using linear, nonlinear, and periodic functions, as well as the nonlinear Henon map, contaminated with different levels of artificial noise for small and large datasets. In this dissertation, 50 and 100 points (comparable to the sizes of the geophysical data sets used in the dissertation) and 1000 points are considered as short and long time series, respectively.

4.4.1 Details of the simulated data

Case 1 (Linear): simple linear functions with Gaussian noise (ε) are used, such as $X \sim N(0,1)$, $Y : y_i = x_i + \varepsilon_i$ for $i = 1, .., N$, where X is independent and identically distributed (iid) and $\varepsilon \sim N(0, \sigma_n)$ is *iid* and independent of X. *Case 2 (Quadratic):* simple quadratic functions with Gaussian noise are used, such as $X \sim N(0,1)$, $Y : y_i = {x_i}^2 + \varepsilon_i$ for $i = 1, .., N$, where X and ε have the same meaning described above. *Case 3 (Periodic):* the periodical system with Gaussian noise is also analyzed, such as X, $Y : y_i = \sin(x_i) + \varepsilon_i$ for $i = 1, .., N$, where X is uniformly distributed between -π to π and $\varepsilon \sim N(0, \sigma_n)$ is *iid* and independent of X. *Case 4 (Chaotic):* the Henon map, which exhibits chaotic behavior, is $H_X : H_{x_{i+1}} = 1 -$

$\alpha {H_{x_i}}^2 + Hy_i$, $H_Y : H_{y_{i+1}} = \beta H_{x_i}$ for $i = 1, .., N$, where $\alpha = 1.4; \beta = 0.3$; and $(H_{x_0}, H_{y_0}) = (0.0, 0.0)$. The Henon map with Gaussian noise is also analyzed, such as $X : x_i = H_{x_i} + \varepsilon x_i$, $Y : y_i = H_{y_i} + \varepsilon y_i$ for $i = 1, .., N$, where $\varepsilon x \sim N(0, \sigma_{H_X})$ and $\varepsilon y \sim N(0, \sigma_{H_Y})$ are *iid* and independent of H_X and H_Y, respectively, and σ_{H_X} and σ_{H_Y} are standard deviations of H_X and H_Y, respectively. The formulations for computing theoretical values of MI for *cases 1, 2, and 3* are described in *section 4.5*.

4.4.2 Conclusion from simulations

The simulations indicate that the presence of noise typically leads to an under-estimation of the *true* MI between the underlying nonlinear signals (see *section 4.5*). As compared to KNN and Edgeworth, KDE is found to capture the underlying nonlinear dependence more consistently between two time series when they are short and noisy assuming such dependence exists (see *section 4.5*). We also compare nonlinear dependence measures, such as KDE, KNN, and Edgeworth, with a rank-based dependence measure, i.e., Kendall's tau. From Kendall's tau, we observe a large negative bias in nonlinear dependence in the simulated data contaminated with noise (see *section 4.5*). Thus in this dissertation LR and KDE approaches have been consistently used to estimate and compare linear and nonlinear CCs.

4.5 Comparisons of MI estimation methods using simulations

4.5.1 Comparison between KDE, KNN, and Edgeworth

The three MI estimation approaches, viz. KDE, KNN, and Edgeworth, are investigated to find the most effective method in terms of quantifying the underlying nonlinear dependence for noisy and short data. Linear and nonlinear dependence measures are evaluated and validated using simulated time series with different linear or nonlinear behavior. The best method should give zero correlation when there is no dependence and quantify any linear or nonlinear correlation which may be present. The presence of noise makes the detection of linear and nonlinear dependence difficult, as this dependence may be obscured if the noise component dominates. This is true especially for short time series and evident from the density plots for linear, nonlinear, periodic, and nonlinear dynamical systems, with different noise to signal ratios (Figures 13-16). With increasing noise levels, the linear and nonlinear dependence cannot be readily discerned visually as shown in the kernel density plots (Figures 13-16). This dissertation chooses the best method which captures more of the *true* nonlinear dependence such that 90% confidence bounds do not overlap with the estimates

for linear correlation (in the case of data known to have nonlinear dependence) even in the presence of noise and for short data sets.

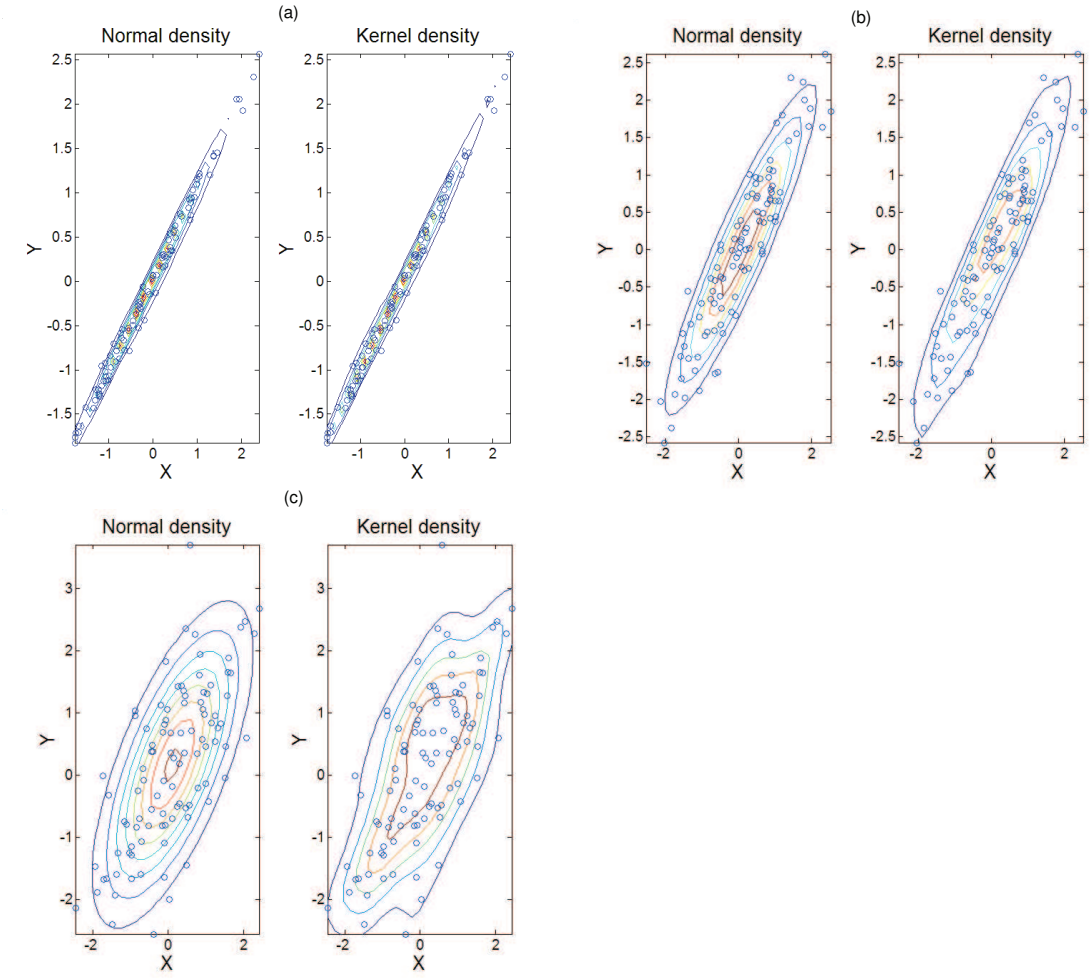

Figure 13. Normal and kernel densities with different noise (σ_n) to signal (σ_s) ratios for *Case 1 (Linear)* with $N = 100$. For kernel density, a Gaussian kernel with optimal Gaussian bandwidth, given as $h = N^{-1/6}$, is used. (a) $\sigma_n/\sigma_s = 0.1$. (b) $\sigma_n/\sigma_s = 0.5$. (c) $\sigma_n/\sigma_s = 1.0$. The linear dependence structure can be seen clearly for cases (a) and (b) but cannot be readily identified for case (c) based on eye estimation.

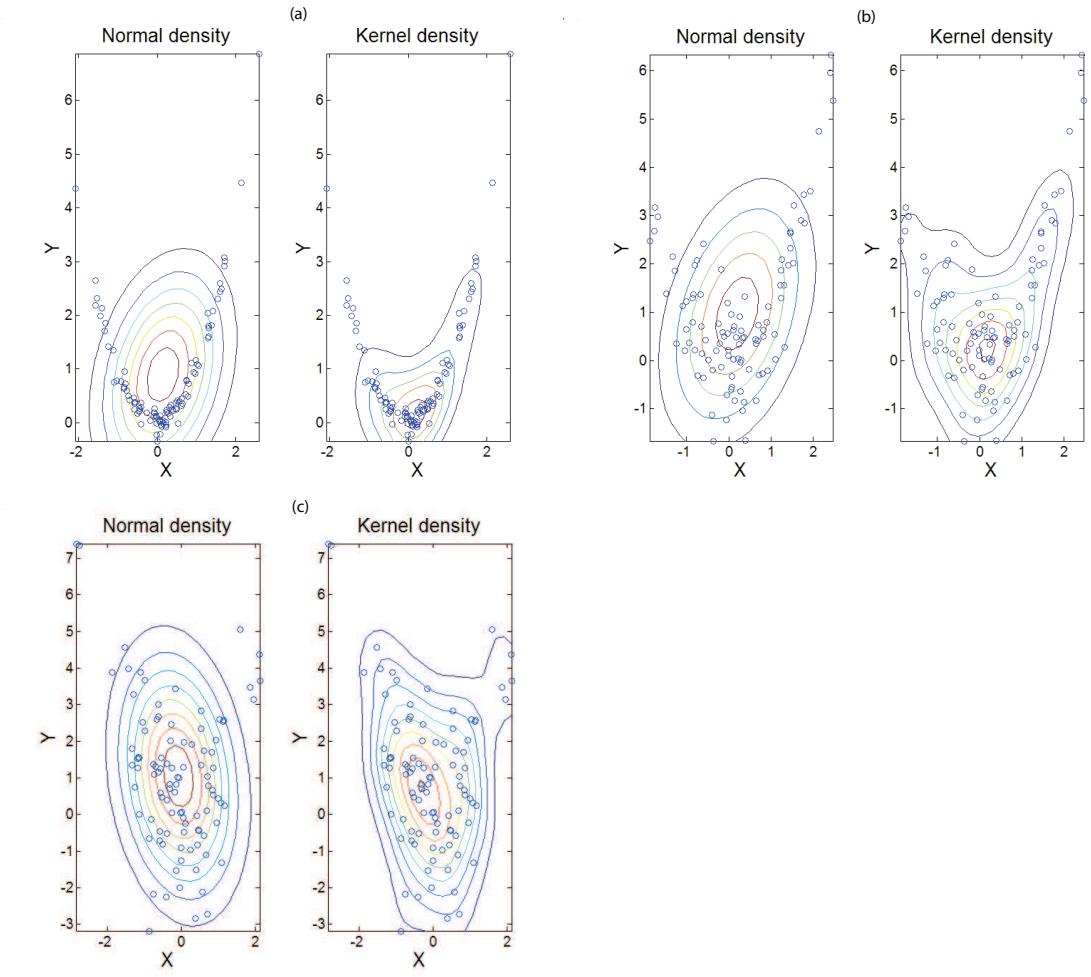

Figure 14. Normal and kernel densities with different noise (σ_n) to signal (σ_s) ratios for *Case 2 (Quadratic)* with $N = 100$. For kernel density, a Gaussian kernel with optimal Gaussian bandwidth, given as $h = N^{-1/6}$, is used. (a) $\sigma_n/\sigma_s = 0.1$. (b) $\sigma_n/\sigma_s = 0.5$. (c) $\sigma_n/\sigma_s = 1.0$. At lower noise levels, such as in cases (a) and (b), the nonlinear dependence can be clearly seen as shown by the kernel density plots. However at higher noise levels, such as in case (c), the dependence structure is not readily discernible visually from the kernel density.

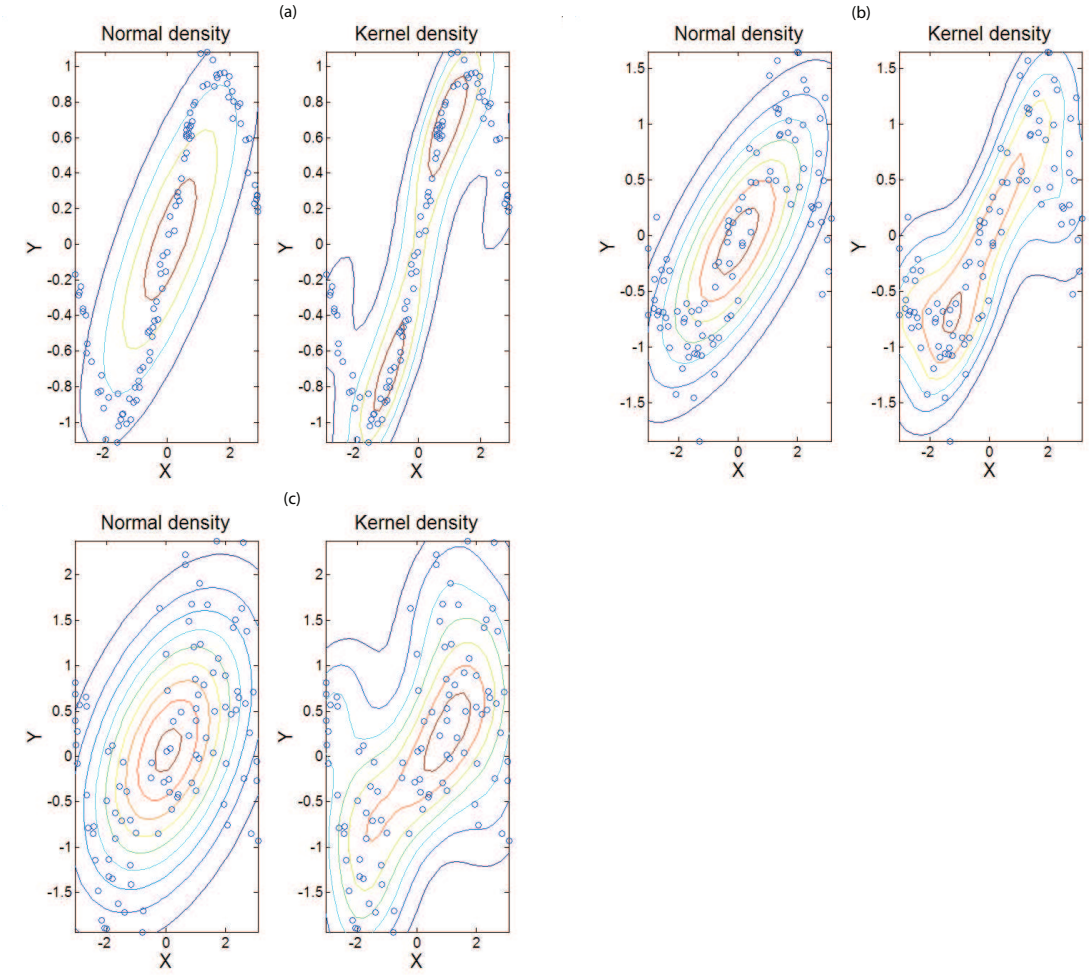

Figure 15. Normal and kernel densities with different noise (σ_n) to signal (σ_s) ratios for *Case 3 (Periodic)* with $N = 100$. For kernel density, a Gaussian kernel with optimal Gaussian bandwidth, given as $h = N^{-1/6}$, is used. (a) $\sigma_n/\sigma_s = 0.1$. (b) $\sigma_n/\sigma_s = 0.5$. (c) $\sigma_n/\sigma_s = 1.0$. With increasing noise levels, the nonlinear dependence structure cannot be identified visually, as shown by the kernel density plots.

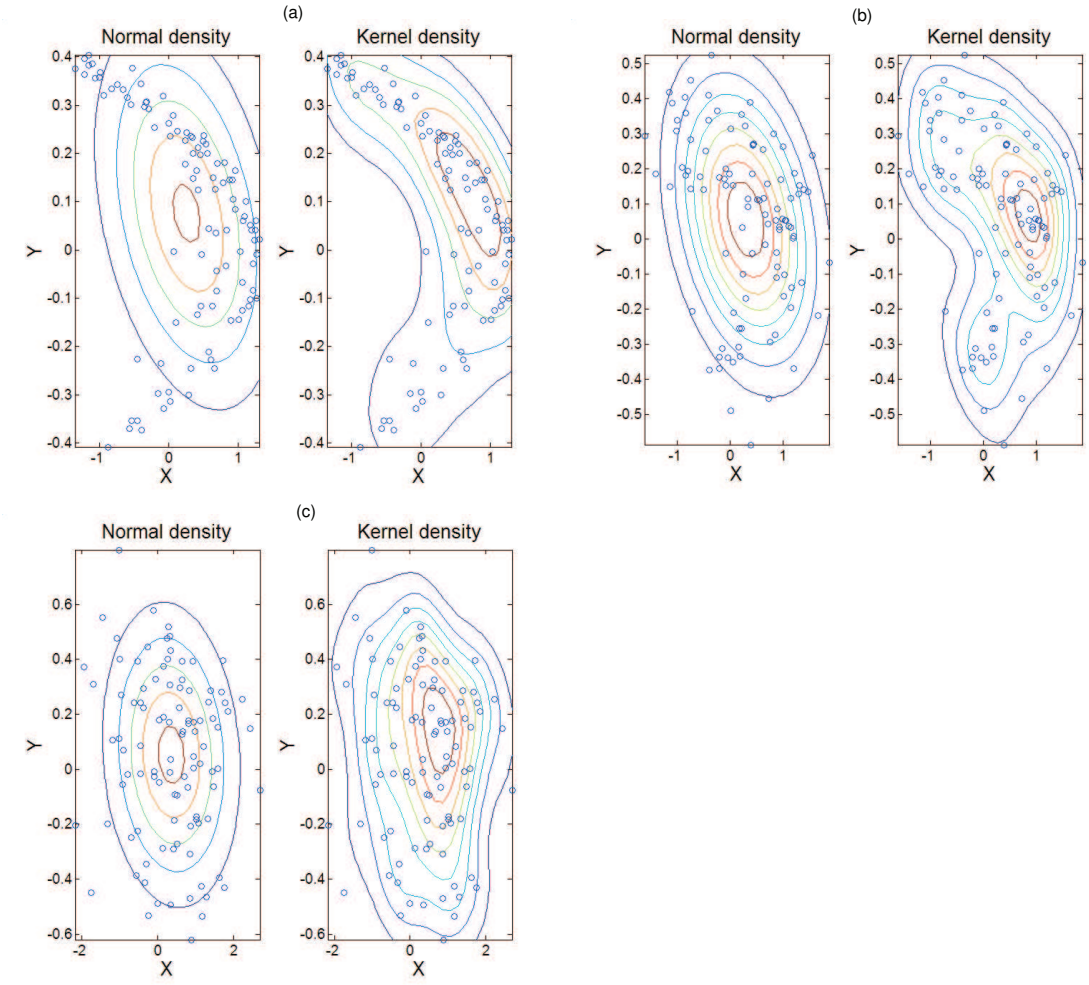

Figure 16. Normal and kernel densities with different noise (σ_n) to signal (σ_s) ratios for *Case 4 (Chaotic)* with $N = 100$. For kernel density, a Gaussian kernel with optimal Gaussian bandwidth, given as $h = N^{-1/6}$, is used. (a) $\sigma_n/\sigma_s = 0.1$. (b) $\sigma_n/\sigma_s = 0.5$. (c) $\sigma_n/\sigma_s = 1.0$. For cases (a) and (b), kernel density plots show the Henon attractor. However the Henon attractor cannot be readily distinguished visually for case (c).

4.5.1.1 Simulation cases and their theoretical values Four examples for simulations namely, linear, short order nonlinear polynomial, periodic (or infinite order polynomial) and nonlinear dynamical, are chosen. In all cases, correlation coefficients (CC) are compared with theoretical CCs which can be computed theoretically.

- *Case 1 (Linear) :* $X \sim N(0,1)$, $Y : y_i = x_i + \varepsilon_i$ for $i = 1,..,N$, where X is independent and identically distributed (iid) and $\varepsilon \sim N(0,\sigma_n)$ is *iid* and independent of X. Let $Z = \varepsilon$, so $Y = X + Z$. The mutual information, I(X;Y), is given as

$$I(X;Y) = H(Y) - H(Y|X), \tag{25}$$

where $H(Y|X) = H(Z) = 0.5\log(2\pi e{\sigma_n}^2)$. The *pdf* of Z is $p_Z(z) = (2\pi)^{-1/2}(\sigma_n)^{-1}\exp(\frac{-z^2}{2{\sigma_n}^2})$. The *pdf* of X is $p_X(x) = (2\pi)^{-1/2}(\sigma_x)^{-1}\exp(\frac{-x^2}{2{\sigma_x}^2})$. In order to compute H(Y), the *pdf* of Y, i.e., $p_Y(y)$, is needed. Since $Y = X + Z$, and X and Z are independent, $p_Y(y)$ is obtained through the convolution of the *pdfs* of X and Z given as

$$p_Y(y) = \int_{-\infty}^{\infty} p_X(x)p_Z(y-x)dx. \tag{26}$$

From equation 26, the *pdf* of Y is given as $p_Y(y) = (2\pi)^{-1/2}({\sigma_x}^2 + {\sigma_n}^2)^{-1/2}\exp[\frac{-x^2}{2({\sigma_x}^2+{\sigma_n}^2)}]$. The entropy of Y is calculated as $H(Y) = \int p_Y(y)\log p_Y(y)dy = 0.5\log[2\pi e({\sigma_x}^2 + {\sigma_n}^2)]$. Substituting $H(Y)$ and $H(Y|X)$ in equation 25, the MI is computed as $I(X;Y) = 0.5\log(1 + \frac{{\sigma_x}^2}{{\sigma_n}^2})$.

- *Case 2 (Quadratic) :* $X \sim N(0,1)$, $Y : y_i = {x_i}^2 + \varepsilon_i$ for $i = 1,..,N$, where X is *iid* and $\varepsilon \sim N(0,\sigma_n)$ is *iid* and independent of X. Let $U = X^2$ and $Z = \varepsilon$, so $Y = U + Z$. The conditional entropy $H(Y|X)$ is given as $H(Y|X) = H(Z) = 0.5\log(2\pi e{\sigma_n}^2)$. The *pdf* of Z is $p_Z(z) = (2\pi)^{-1/2}(\sigma_n)^{-1}\exp(\frac{-z^2}{2{\sigma_n}^2})$. The *pdf* of U is $p_U(u) = \begin{cases} (2\pi)^{-1/2}(u)^{-1/2}\exp(\frac{-u}{2}), & u > 0 \\ 0, & otherwise. \end{cases}$
In order to compute H(Y), the *pdf* of Y, i.e., $p_Y(y)$, is needed. Since $Y = U + Z$, and U and Z are independent, $p_Y(y)$ is obtained through the convolution of the *pdfs* of U and Z given as

$$p_Y(y) = \int_{-\infty}^{\infty} p_U(u)p_Z(y-u)du. \tag{27}$$

Equation 27 is solved using numerical integration for different values of σ_n. $H(Y)$, given as $H(Y) = \int p_Y(y) \log p_Y(y) dy$, is computed using $p_Y(y)$. The MI is computed by substituting $H(Y)$ and $H(Y|X)$ in equation 25.

- *Case 3 (Periodic) :* X, $Y : y_i = \sin(x_i) + \varepsilon_i$ for $i = 1, .., N$, where X is uniformly distributed between -π to π and $\varepsilon \sim N(0, \sigma_n)$ is *iid* and independent of X. Let $V = \sin(X)$ and $Z = \varepsilon$, so $Y = V + Z$. The conditional entropy $H(Y|X)$ is given as $H(Y|X) = H(Z) = 0.5 \log(2\pi e {\sigma_n}^2)$. The *pdf* of Z is $p_Z(z) = (2\pi)^{-1/2} (\sigma_n)^{-1} \exp(\frac{-z^2}{2{\sigma_n}^2})$. The *pdf* of V is $p_U(u) = (\pi)^{-1}(1 - v^2)^{-1/2}$ for $0 \leq v < 1$. In order to compute H(Y), the *pdf* of Y, i.e., $p_Y(y)$, is needed. Since $Y = V + Z$, and V and Z are independent, $p_Y(y)$ is obtained through the convolution of the *pdfs* of V and Z given as

$$p_Y(y) = \int_{-\infty}^{\infty} p_V(v) p_Z(y - v) dv. \tag{28}$$

Equation 28 is solved using numerical integration for different values of σ_n. $H(Y)$, given as $H(Y) = \int p_Y(y) \log p_Y(y) dy$, is computed using $p_Y(y)$. The MI is computed by substituting $H(Y)$ and $H(Y|X)$ in equation 25.

- *Case 4 (Chaotic) :* the Henon map, which exhibits chaotic behavior, is $H_X : H_{x_{i+1}} = 1 - \alpha {H_{x_i}}^2 + H y_i$, $H_Y : H_{y_{i+1}} = \beta H_{x_i}$ for $i = 1, .., N$, where $\alpha = 1.4; \beta = 0.3$; and $(H_{x_0}, H_{y_0}) = (0.0, 0.0)$. The Henon map with Gaussian noise is analyzed, such as $X : x_i = H_{x_i} + \varepsilon x_i$, $Y : y_i = H_{y_i} + \varepsilon y_i$ for $i = 1, .., N$, where $\varepsilon x \sim N(0, \sigma_{H_X})$ and $\varepsilon y \sim N(0, \sigma_{H_Y})$ are *iid* and independent of H_X and H_Y, respectively, and σ_{H_X} and σ_{H_Y} are standard deviations of H_X and H_Y, respectively. For the nonlinear dynamical example, theoretical values of $I(X;Y)$ could not be computed and were not found anywhere in the literature, neither for the specific example of Henon map nor for any other nonlinear dynamical time series.

In each case three sets of data of size (N) 50, 100, and 1000 points are used. In this dissertation, 50 and 100 points (comparable to the sizes of the geophysical data sets used in the dissertation) and 1000 points are considered as short and long time series, respectively. The total number of samples considered for $N = 50$, $N = 100$, and $N = 1000$ are 200, 100, and 20, respectively. CCs are defined as the mean CC from the total number of samples. The 90% confidence bounds of CCs are given by 5% and 95% quantiles of CCs obtained from the total samples. The correlation coefficients are plotted against noise (σ_n) to signal (σ_s) ratio, i.e. σ_n / σ_s. CCs obtained from KDE, KNN, Edgeworth, and LR are compared to find the best method which

consistently captures the *true* nonlinear dependence given by theoretical CCs and its 90% confidence bounds do not overlap with the bounds from LR for short and noisy data sets.

Table 5. MI estimates with standard errors given in parentheses between two Gaussian noise sets (X_i, Y_i) : $X \sim N(0,1), Y \sim N(0,1), i = 1,..,N$, where X and Y are *iid* and independent of each other. The total number of samples for $N = 50$, $N = 100$, and $N = 1000$ are 200, 100, and 20, respectively. The MI estimates and its standard errors are the mean and standard deviation from the total samples. The MI should be zero between two Gaussian noise sets. The MI estimates obtained from all three methods are close to zero but biased upwards in the case of KDE and KNN.

Method	50	100	1000
KDE	0.1033 (0.0351)	0.0707 (0.0192)	0.0270 (0.0046)
KNN	0.0829 (0.0549)	0.0555 (0.0425)	0.0220 (0.0143)
Edgeworth	0.0298 (0.0282)	0.0139 (0.0115)	0.0016 (0.0012)

The best method needs to be robust to noise and short data, both in the sense that the dependence among the underlying nonlinearities are captured as well as the computed dependence is indeed zero when the variables are known to be independent. The MI between two independent variables is zero and this fact can be utilized as a *consistency check* for each method. Here two Gaussian noise sets having zero mean and unit variance are analyzed. The MI estimates are found to be close to zero from all three methods (Table 5). This demonstrates that all three methods, viz. KDE, KNN, and Edgeworth, pass the consistency test and yield zero dependence when the two variables are independent. Edgeworth gives the best estimates and smallest error bounds between two uncorrelated Gaussians. The MI estimates from KDE and KNN are biased upwards. KDE is preferable over KNN as the standard errors from KNN are more than 1.5 times larger than those from KDE.

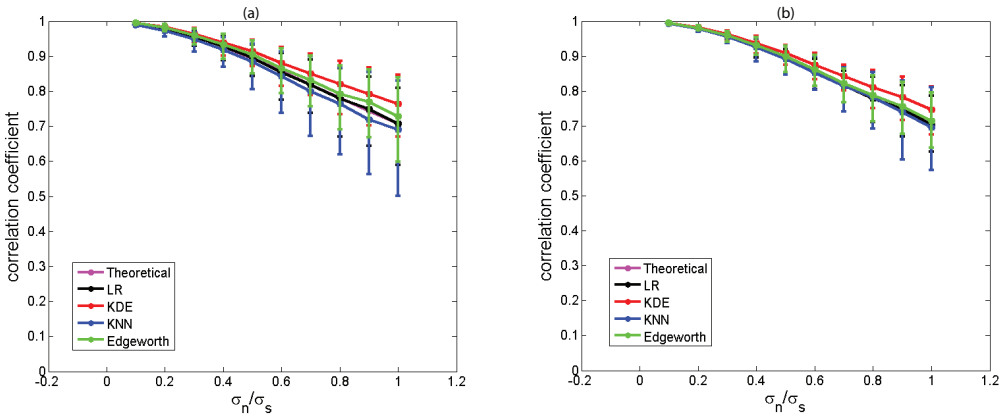

Figure 17. Nonlinear and linear CCs for *Case 1 (Linear)* with 90% confidence bounds obtained from KDE and LR, respectively. (a) $N = 50$. (b) $N = 100$. In all cases, linear and nonlinear estimates from all three methods overlap with theoretical CCs indicating that the linear and nonlinear estimation methods capture the true dependence when there is only a linear dependence. But at higher noise levels, KDE seems to have an edge over KNN and Edgeworth because of its narrow bounds.

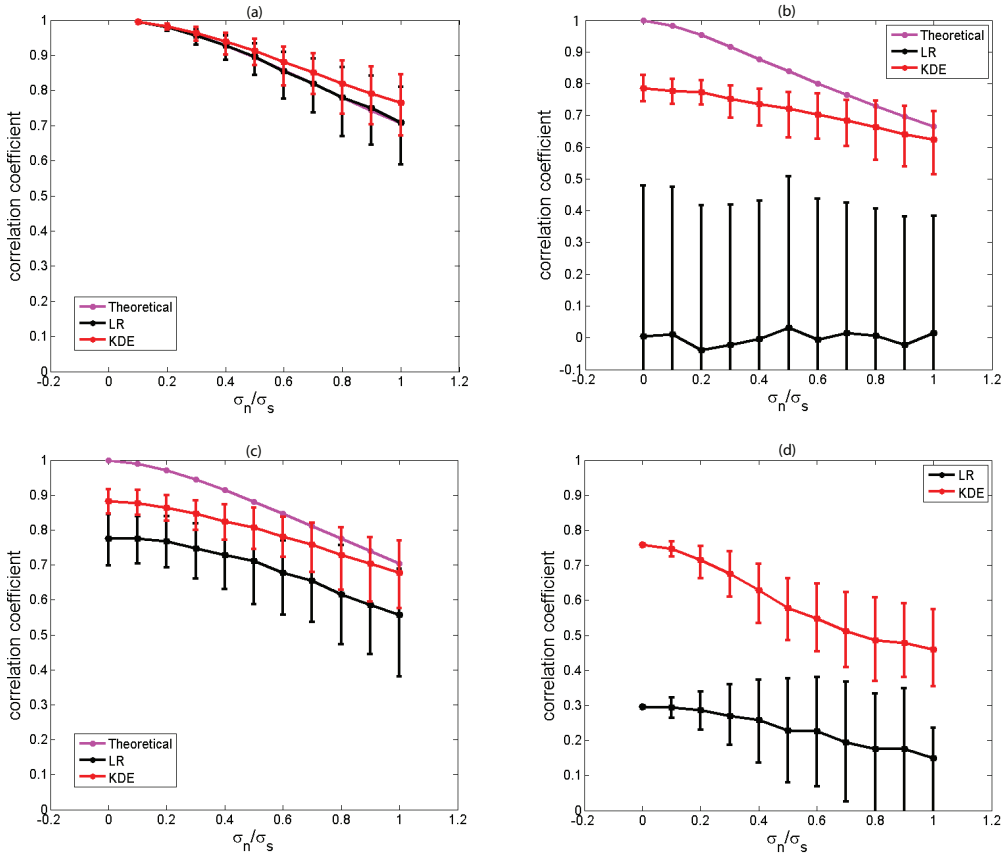

Figure 18. Nonlinear and linear CCs between functions, such as (a) Case 1 (Linear); (b) Case 2 (Quadratic); (c) Case 3 (Periodic); and (d) Case 4 (Chaotic), and their 90% confidence bounds are obtained using KDE and LR, respectively. CCs and their 90% bounds are obtained from 200 samples of size $N = 50$. At higher noise levels, KDE captures the true dependence given by theoretical CCs as shown in (a), (b), and (c). In (c), KDE estimates are not different from linear CCs considering 90% confidence bounds. In (b) and (d), KDE gives more correlation as compared to the linear correlation and there is a clear separation between their 90% confidence bounds.

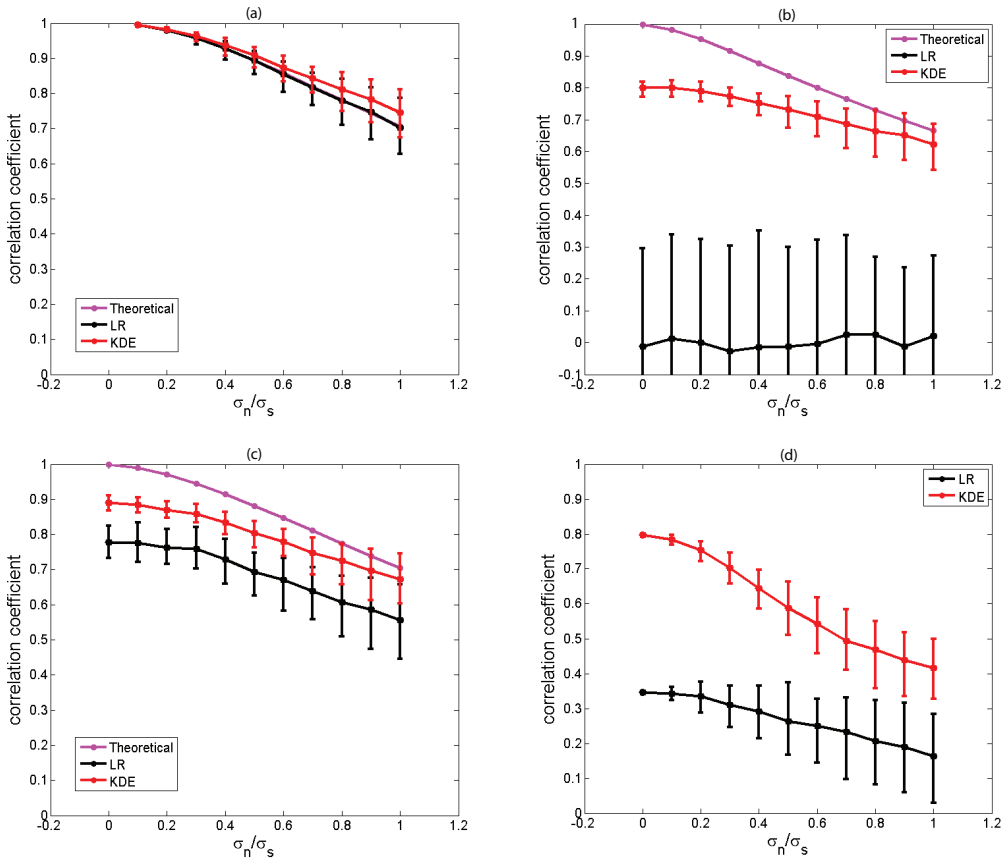

Figure 19. Nonlinear and linear CCs between functions, such as (a) Case 1 (Linear); (b) Case 2 (Quadratic); (c) Case 3 (Periodic); and (d) Case 4 (Chaotic), and their 90% confidence bounds are obtained using KDE and LR, respectively. CCs and their 90% bounds are obtained from 100 samples of size $N = 100$. At higher noise levels, KDE captures the true dependence given by theoretical CCs as shown in (a), (b), and (c). In (c), KDE estimates are not different from linear CCs considering 90% confidence bounds. In (b) and (d), KDE gives more correlation as compared to the linear correlation and there is a clear separation between their 90% confidence bounds.

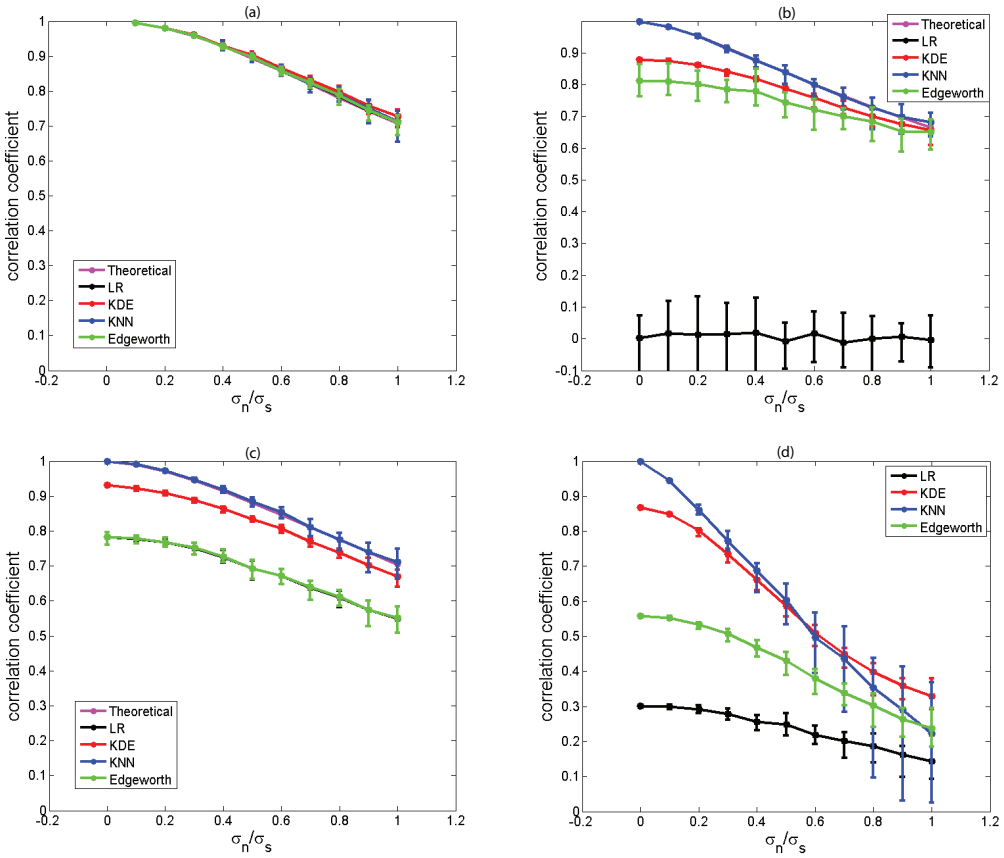

Figure 20. Nonlinear and linear CCs with 90% confidence bounds obtained from KDE and LR, respectively, using $N = 1000$ points . (a) Case 1 (Linear); (b) Case 2 (Quadratic); (c) Case 3 (Periodic); and (d) Case 4 (Chaotic). At lower noise levels, KNN seems to the best as it overlaps with theoretical CCs and has narrow bounds. In (c), linear and Edgeworth estimates overlap exactly. The performance of Edgeworth is not good in (c) and (d). At higher noise levels, KDE and KNN estimates overlap and also include theoretical CCs but KNN estimates also overlap with linear CCs in (d). Thus, KDE seems to have an edge over KNN as its 90% confidence bounds are narrow and do not overlap with linear CCs when the data is noisy and relatively large.

4.5.1.2 Short time series Case 1 (Linear): Analysis with increasing levels of noise shows that CCs decrease with increasing noise to signal ratios. This is expected as the noise component obscures the underlying linear dependence. Nonlinear CCs obtained from all three methods are close and overlap with theoretical values (which, in turn, are expected to be identical to linear CCs in this case), indicating all the approaches (Linear, KDE, KNN, Edgeworth) succeed in capturing the dependence between the random variables (Figures 17, 19a, and 18a). Nonlinear CCs from KDE are biased upwards for several noise to signal ratios (Figures 19a, 18a, and 20a). At higher noise levels, confidence bounds obtained from KNN and Edgeworth are much wider as compared to KDE's confidence bounds (Figures 17 and 20a). While all correlation methods capture the true linear dependence, KDE is the better choice as it has narrower confidence bounds.

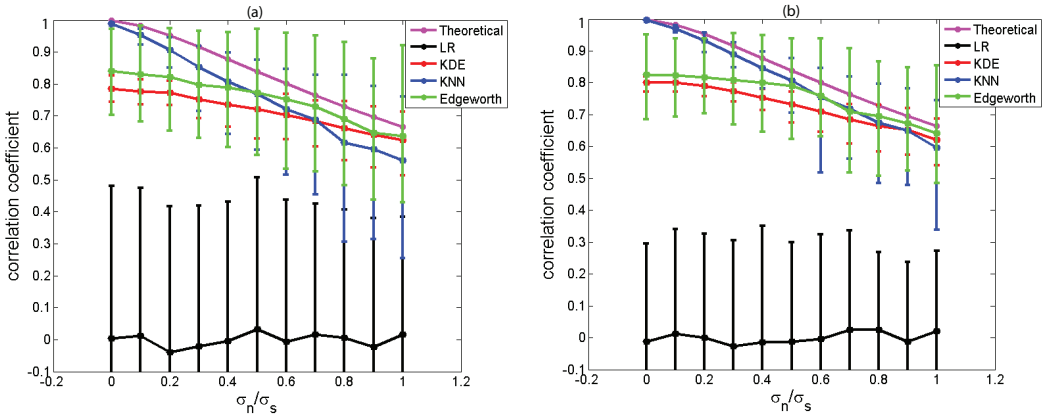

Figure 21. Nonlinear and linear CCs for *Case 2 (Quadratic)* with 90% confidence bounds obtained from KDE and LR, respectively. (a) $N = 50$. (b) $N = 100$. All three nonlinear correlation estimates include theoretical CCs but 90% confidence bounds from KNN and Edgeworth also overlap with linear CCs at higher noise levels in (a). This shows that KNN and Edgeworth estimates are not different from linear CCs at higher noise levels. KDE quantifies more correlation as compared to the linear correlation as their 90% confidence bounds do not overlap indicating that KDE can truly capture the nonlinear dependence.

Case 2 (Quadratic): At all noise levels, linear CCs are close to zero for all values of N (Figures 21 and 20b). Estimates of nonlinear CCs obtained from KNN and Edgeworth include theoretical CCs within their 90% confidence bounds but do not overlap with linear CCs when the noise level is low (Figure 21). But at higher noise levels, confidence bounds from KNN and Edgeworth do overlap with linear CCs, while there is a significant and marked difference between KDE and LR (Figures 19b, 18b, and 21). Thus, at higher noise levels, the performance of KNN and Edgeworth deteriorates because of their large error bounds, making them difficult to use. On the contrary, although KDE underestimates nonlinear CCs for low noise levels it slowly approaches theoretical values at high noise levels with narrow confidence bounds (Figures 21 and 20b). This indicates that KDE may be the better choice for estimating nonlinear dependence when the noise level is high. For lower noise levels KDE can be further improved by choosing a smaller value of the smoothing parameter involved in the computation.

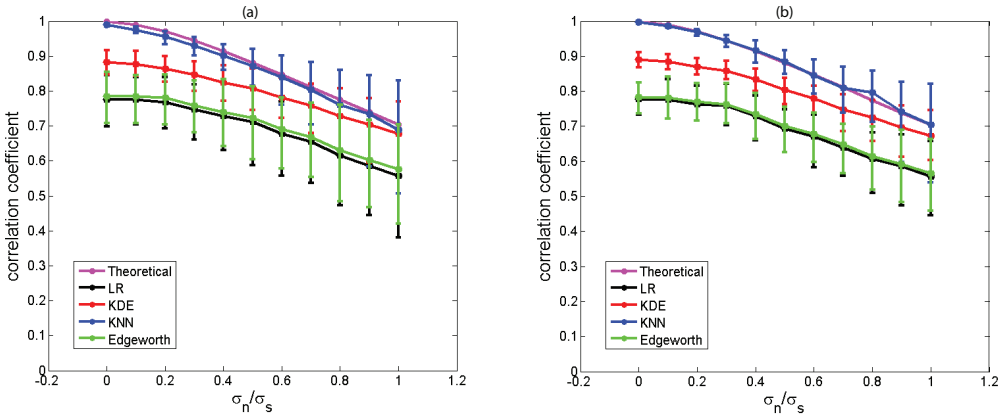

Figure 22. Nonlinear and linear CCs for *Case 3 (Periodic)* with 90% confidence bounds obtained from KDE and LR, respectively. (a) $N = 50$. (b) $N = 100$. Edgeworth captures nothing more than the linear correlation. At low noise levels, KNN seems to the best as it overlaps with theoretical CCs and its bounds are narrow. At higher noise levels, KDE and KNN CCs overlap and also include linear and theoretical CCs but 90% confidence bounds from KNN are wider than that obtained from KDE. At higher noise levels and for relatively small data, KDE seems to have an edge over KNN because of its narrow bounds.

Case 3 (Periodic): KNN seems to be the best in capturing the true nonlinear dependence for any data size and noise levels except for the fact that at higher noise levels it produces wider confidence bounds (Figures 19c, 18c, 22, and 20c). Edgeworth appears to capture only the linear correlation and produces wider confidence bounds (Figures 22 and 20c). In this example the density of Y is bimodal, which causes Edgeworth estimates to be incorrect. Nonlinear CCs from KDE are biased downwards for low noise levels, but they slowly approach theoretical values at high noise levels with narrow confidence bounds (Figures 22 and 20c). Again, KNN is better than KDE at lower noise levels but both KDE and KNN produce better results than Edgeworth, as far as capturing the true dependence is concerned at higher noise levels (Figure 22). For higher noise levels, confidence bounds obtained from KDE, KNN, and Edgeworth overlap with theoretical and linear CCs indicating all these methods capture nothing more than the linear correlation but KDE does seem to have an edge over others as it produces narrower confidence bounds (Figure 22). Thus, KDE appears to be a better choice at higher noise levels.

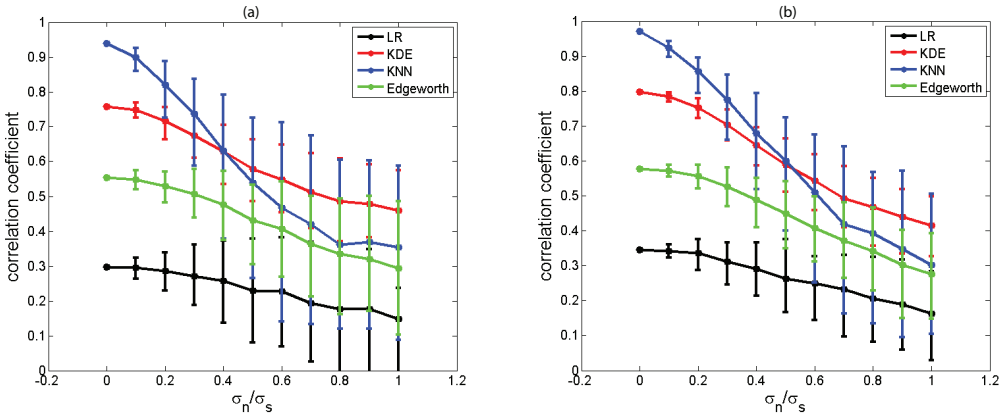

Figure 23. Nonlinear and linear CCs for *Case 4 (Chaotic)* with 90% confidence bounds obtained from KDE and LR, respectively. (a) $N = 50$. (b) $N = 100$. At higher noise levels, KNN and Egdeworth CCs overlap with linear CCs indicating that they do not capture anything more than the linear correlation. KDE is the best in capturing the nonlinear dependence as its 90% confidence bounds do not overlap with linear CCs.

Case 4 (Chaotic): At all noise levels, linear CCs between X and Y components of the Henon map are negative. Since nonlinear CCs from KDE, KNN, and Edgeworth do not have directionality, the absolute values of linear CCs are considered and plotted. Analyses using short time series also show that confidence bounds for CCs from KDE and LR are well separated while confidence bounds obtained from KNN and Edgeworth do overlap with linear CCs at higher noise levels (Figures 19d, 18d, and 23). Note that theoretical CCs for the Henon map could not be computed and were not found in the literature. At lower noise levels, KDE and KNN perform better but at higher noise levels KDE outperforms KNN and Edgeworth. In fact, KDE stands out by its ability to capture more correlation than purely linear correlation (Figure 23).

The Henon example above is yet another case where at higher noise levels, for known nonlinear dependence between the variables, estimates of CCs based on KNN and Edgeworth, because of their large error bounds, cannot conclusively show the presence of that dependence. Although their performance is often better than KDE for small noise levels, they fail in large noise scenarios in small datasets. Since the data is short and highly noisy in geophysics, this simulation exercise points at KDE as the best estimator of MI/CCs among the estimators available.

4.5.1.3 Long time series Analyses of 1000 points show that both KDE and KNN perform better than the others in that they are able to capture more dependence than linear correlation. KNN appears to capture the true dependence better at lower noise levels (Figures 20a, 20b, and 20c). But at higher noise levels all the three nonlinear correlation methods perform equally well only for *case 1* and *case 2* (Figures 20a and 20b). Edgeworth captures nothing more than the linear correlation in *case 3* for all noise levels (Figure 20c). At higher noise levels, the 90% confidence bounds from KNN and Edgeworth overlap with linear CCs for *case 4* (Figure 20d). KDE appears to be the better choice at higher noise levels because its 90% confidence bounds are narrow and do not overlap with linear CCs for all cases. Thus, KNN is the best at lower noise levels whereas KDE is recommended for higher noise levels.

4.5.1.4 Conclusion from the analysis of simulations The accuracy of LR and KDE are estimated based on correspondence to theoretical dependence where these are available or can be computed, which provides a measure of bias, as well as the performance of the confidence bounds, which provides a measure of variance. In addition, the expected lower bounds on predictability can be computed in terms of an error statistic (MSE). The analysis described in the previous sections designate the KDE approach as a better quantifier of the dependence between time series, especially when the number of data points are small and subject to noise.

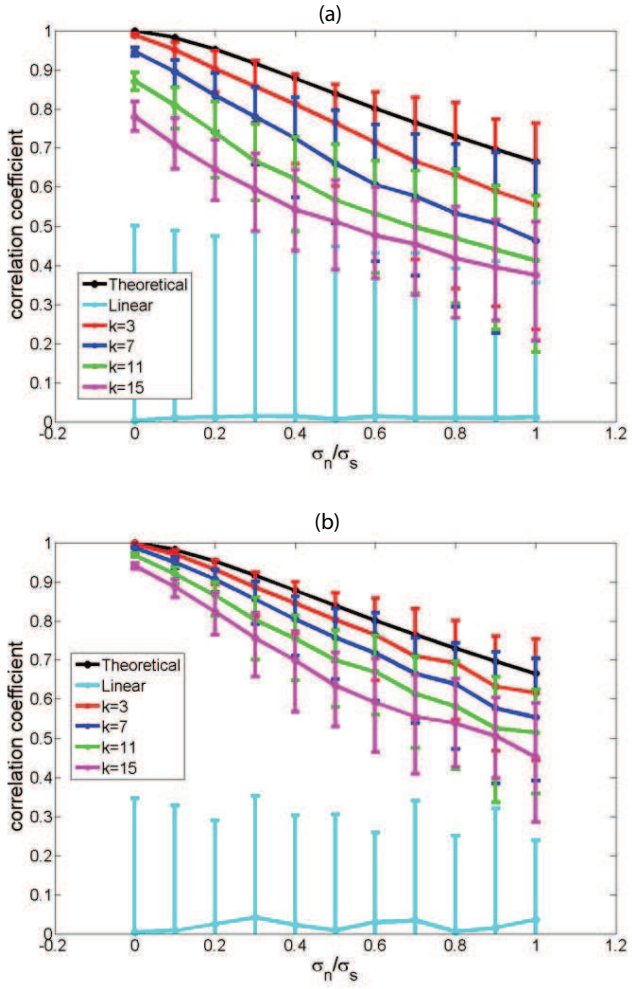

Figure 24. Comparison of correlation coefficients obtained for *Case 2 (Quadratic)* from KNN with different number of nearest neighbors (k), i.e., 3, 7, 11, and 15. (a) $N = 50$. (b) $N = 100$. As k increases, both the bias and variance increase at lower noise levels. For higher noise levels, the bias increases but the variance decreases as k increases.

4.5.1.5 Discussion from the analysis of simulations Our results from the simulated data indicate that KNN performs better in low noise and KDE in high noise situations. An examination of the density plots may explain the relative performance of KDE and KNN (Figures 13-16). The relative performance of KDE and KNN with respect to noise levels is possibly a consequence of the bias-variance tradeoff, since prior literature suggests that the KDE estimate can often be highly biased if the particular KDE recipe used here is followed, while KNN estimates can have significant variance when the number of nearest neighbors (k) is set to low values, e.g., three as used in this dissertation. Thus for low noise to signal ratios, the bias in the KDE estimate dominates the variance of estimate. Thus, the KNN does relatively better for low noise since the bias is lower than KDE and the variance is small as a result of the lower noise levels. However, for high noise to signal ratio, the converse is true, and hence the KDE performs relatively better. In this case variance dominates because of the noise in the data but the variance associated with $k = 3$ for KNN increases dramatically. One way to address the large variance from KNN is to use a much larger value of k but it would also increase the bias (Figure 24). In general, these discussions suggest that the results for nonlinear dependence obtained from KDE and KNN may actually reflect the lower bounds of what may be possible if an improvement or combination of KNN and KDE are used. Specifically, both the KDE and KNN estimates can potentially improve dramatically by utilizing a plug-in method for kernel, bandwidth or k selection. Such plug-in procedures would cause additional estimation variance but may reduce the overall MSE of estimation. However, the development or utilization of procedures for the selection of optimal kernels, bandwidths or nearest neighbors may be rather involved and hence is left as an area of future research.

4.5.2 Comparison of nonlinear dependence measures with a rank-based dependence

Kendall's tau is a rank-based dependence measure used to estimate the strength of dependence between variables. It provides a robust approach to assess monotonic nonlinear dependence. In fact, linear and rank correlation have been used jointly to understand the nature of dependence in the literature [100]. Let $(X, Y) \sim F$ be a pair of random variables with distribution F and $(\tilde{X}, \tilde{Y}) \sim F$ be independent of (X, Y) (and with the same distribution F), Kendall's tau between X and Y is defined as

$$\begin{aligned} \tau = \quad & P\left(\left(X - \tilde{X}\right)\left(Y - \tilde{Y}\right) > 0\right) \\ & -P\left(\left(X - \tilde{X}\right)\left(Y - \tilde{Y}\right) < 0\right). \end{aligned} \tag{29}$$

The empirical estimator of Kendall's tau for an *iid* sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ is

$$\hat{\tau} = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \text{sign}\left((X_i - X_j)(Y_i - Y_j)\right). \tag{30}$$

The complete description of Kendall's tau is given in [101].

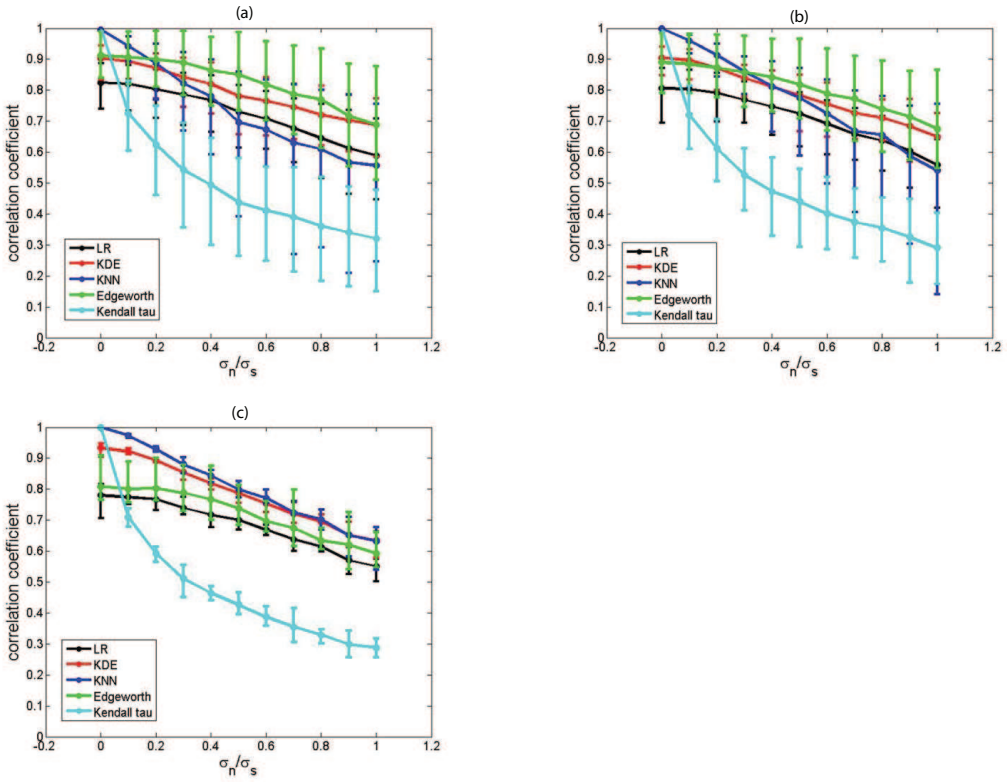

Figure 25. Correlation coefficients for *Case - Cubic* with 90% confidence bounds obtained from LR, KDE, KNN, Edgeworth, and Kendall's tau. (a) $N = 50$. (b) $N = 100$. (c) $N = 1000$. For all cases, the lowest curve is obtained from Kendall's tau. In (a) and (b), Kendall's tau overlaps with linear at lower noise levels. But at higher noise levels, it overlaps with KNN. In (c), Kendall's tau captures the lowest dependence.

The performance of Kendall's tau is compared with KDE, KNN, and Edgeworth using the simulated data. Since Kendall's tau is used to assess monotonic nonlinear dependence, it would not work for the test cases described in *section 1.1*, i.e., quadratic, periodic, and chaotic. Therefore, we use a different test case given as *Case - Cubic :* $X \sim N(0,1),\ Y : y_i = {x_i}^3 + \varepsilon_i$ for $i = 1,..,N$, where X is *iid* and $\varepsilon \sim N(0,\sigma_n)$ is *iid* and independent of X. Since we are just interested in comparing Kendall's tau with nonlinear estimates from KDE, KNN, and Edgeworth, we do not compute theoretical estimates for this case. We consider three different number of points, i.e., 50 and 100 (short series), and 1000 (long series). For all three cases, Kendall's tau is one when there is no noise (Figure 25). The variance of Kendall's tau increases as the number of points decreases. As the noise levels increase, nonlinear dependence from Kendall's tau deteriorates and falls below linear (LR), KDE, KNN, and Edgeworth. For short series, Kendall's tau overlaps with linear at lower noise levels indicating that it captures nothing more than the linear dependence (Figures 25a and 25b). At higher noise levels, Kendall's tau overlaps with KNN for short series where KNN does not perfom well and produces large variances. Therefore, Kendall's tau is not a good estimator for short series. For long series, Kendall's tau produces the lowest dependence and its 90% confidence limits do not overlap with any estimator (Figure 25c). It shows that it cannot be utilized for long series also. These simulations emphasize one important point that the performance of Kendall's tau deteriorates even at a presence of small amount of noise.

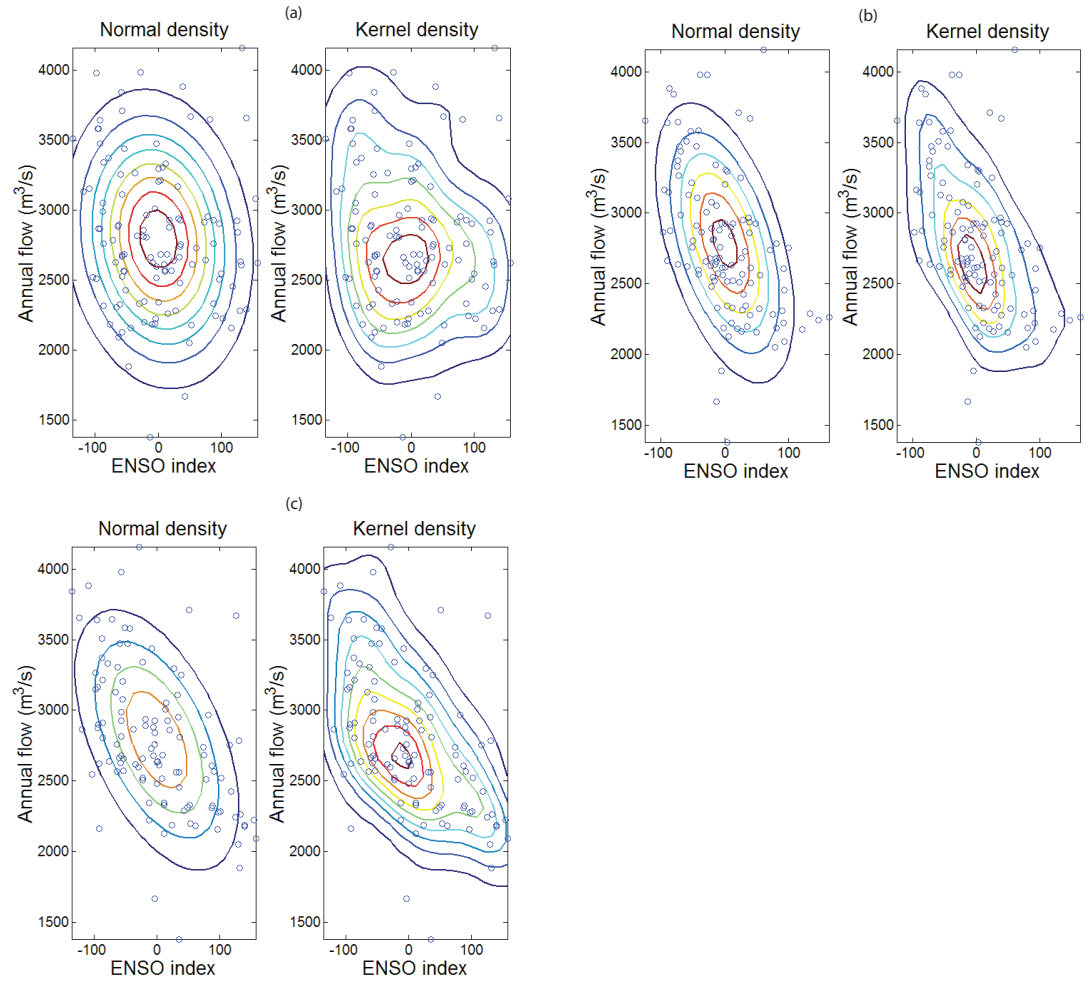

Figure 26. The bivariate normal and kernel density between the ENSO index for different quarters and the annual flow of the Nile River. For kernel density, a Gaussian kernel with optimal Gaussian bandwidth, given as $h = N^{-1/6}$ where N is the total number of observations, is used. (a) Quarter 1. (b) Quarter 4. (c) Quarter 5. Quarter 1 and 5 show the lowest and highest linear CCs between the ENSO index and the Nile flow, respectively (Table 6). Quarter 1 and 4 show the lowest and highest nonlinear CCs between the ENSO index and the Nile flow, respectively (Table 6).

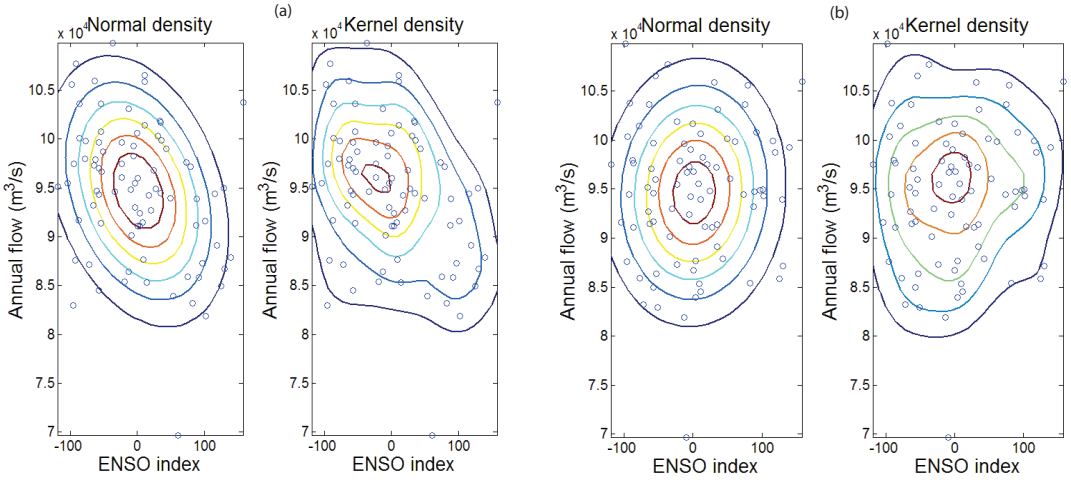

Figure 27. The bivariate normal and kernel density between the ENSO index for different quarters and the annual flow of the Amazon River. For kernel density, a Gaussian kernel with optimal Gaussian bandwidth, given as $h = N^{-1/6}$ where N is the total number of observations, is used. (a) Quarter 3. (b) Quarter 7. Quarter 7 and 3 show the lowest and highest linear and nonlinear CCs between the ENSO index and the Amazon flow, respectively (Table 7).

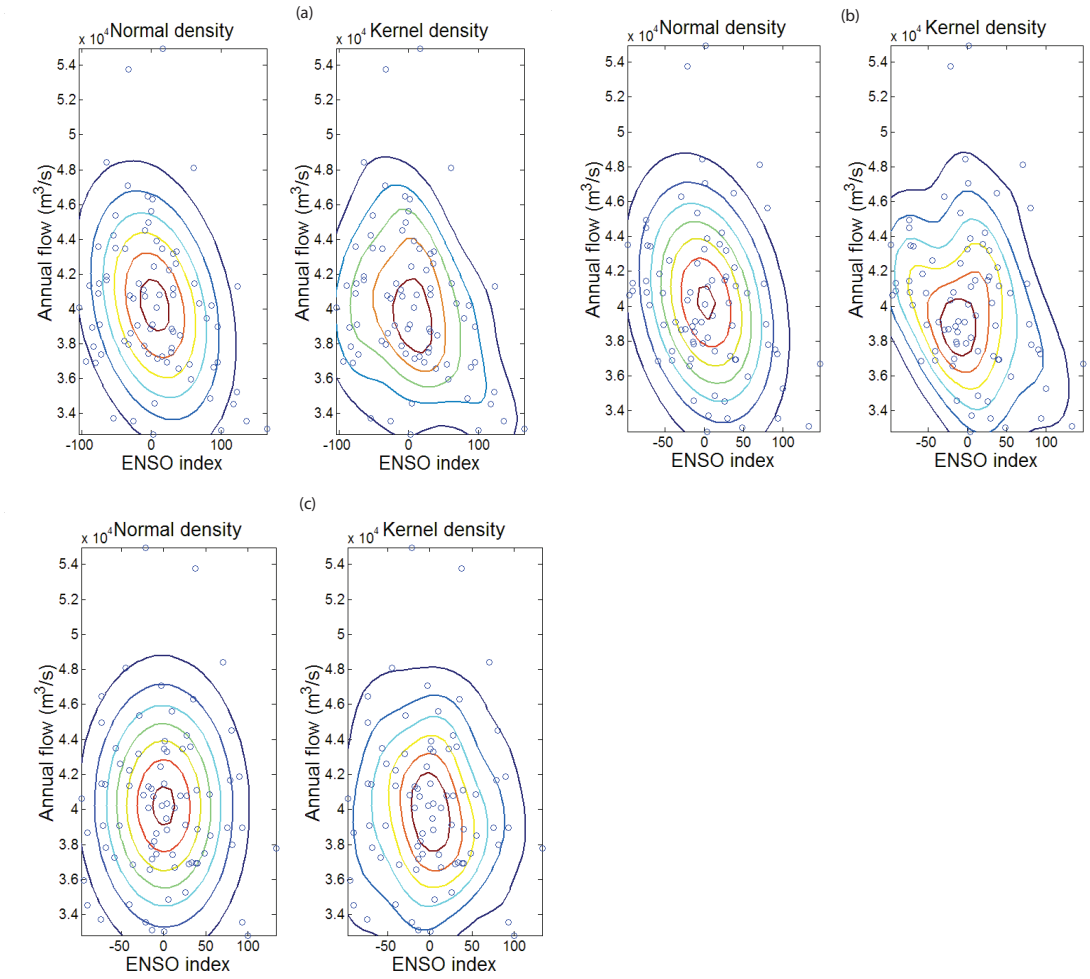

Figure 28. The bivariate normal and kernel density between the ENSO index for different quarters and the annual flow of the Congo River. For kernel density, a Gaussian kernel with optimal Gaussian bandwidth, given as $h = N^{-1/6}$ where N is the total number of observations, is used. (a) Quarter 2. (b) Quarter 3. (c) Quarter 7. Quarter 7 and 2 show the lowest and highest linear CCs between the ENSO index and the Congo flow, respectively (Table 8). Quarter 7 and 3 shows the lowest and highest nonlinear CCs between the ENSO index and the Congo flow, respectively (Table 8).

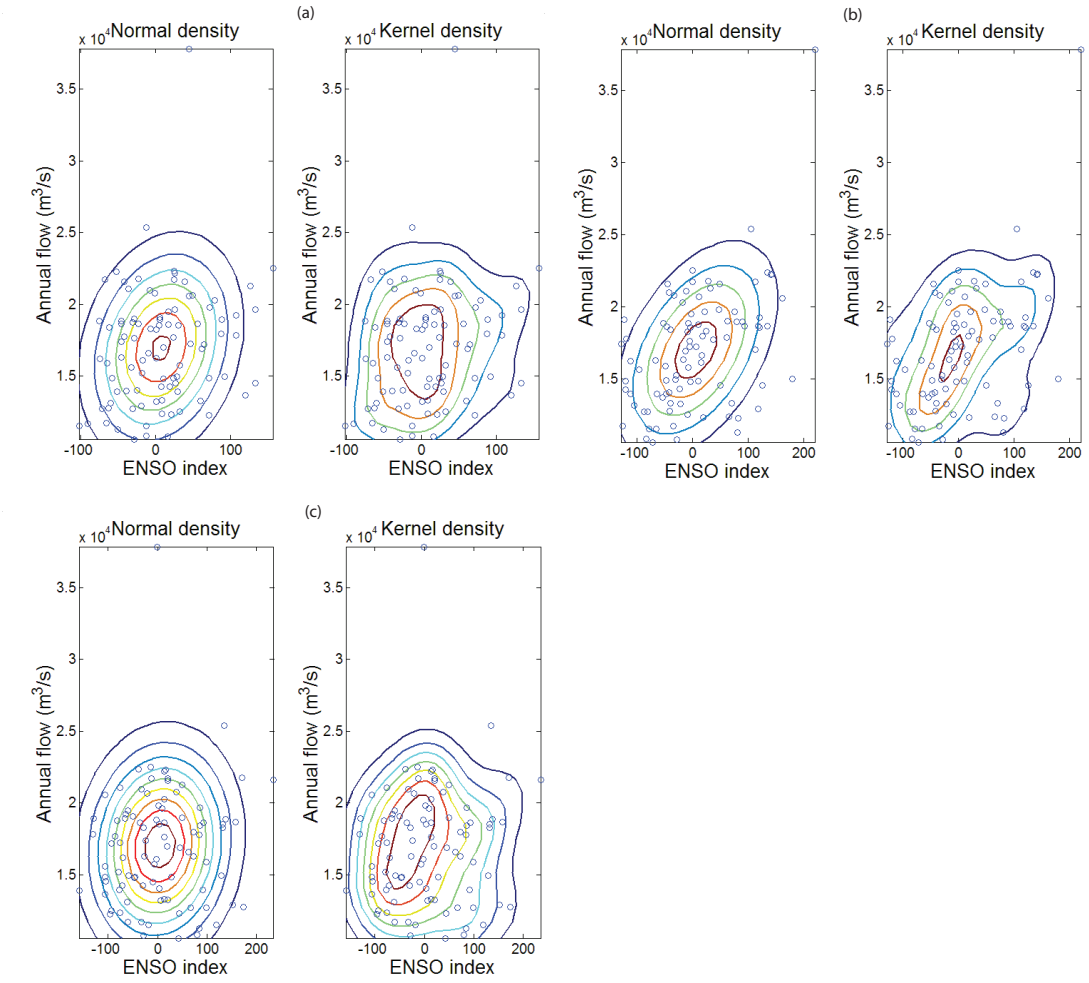

Figure 29. The bivariate normal and kernel density between the ENSO index for different quarters and the annual flow of the Paraná River. For kernel density, a Gaussian kernel with optimal Gaussian bandwidth, given as $h = N^{-1/6}$ where N is the total number of observations, is used. (a) Quarter 2. (b) Quarter 5. (C) Quarter 8. Quarter 8 and 5 show the lowest and highest linear CCs between the ENSO index and the Paraná flow, respectively (Table 9). Quarter 2 and 5 show the lowest and highest nonlinear CCs between the ENSO index and the Paraná flow, respectively (Table 9).

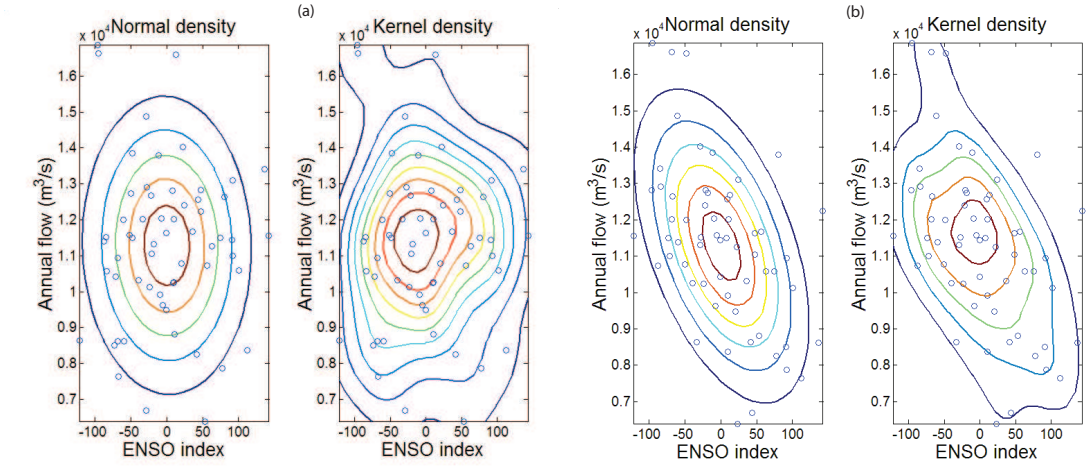

Figure 30. The bivariate normal and kernel density between the ENSO index for different quarters and the annual flow of the Ganges River. For kernel density, a Gaussian kernel with optimal Gaussian bandwidth, given as $h = N^{-1/6}$ where N is the total number of observations, is used. (a) Quarter 1. (b) Quarter 5. Quarter 1 and 5 shows the lowest and highest linear and nonlinear CCs between the ENSO index and the Ganges flow, respectively (Table 10).

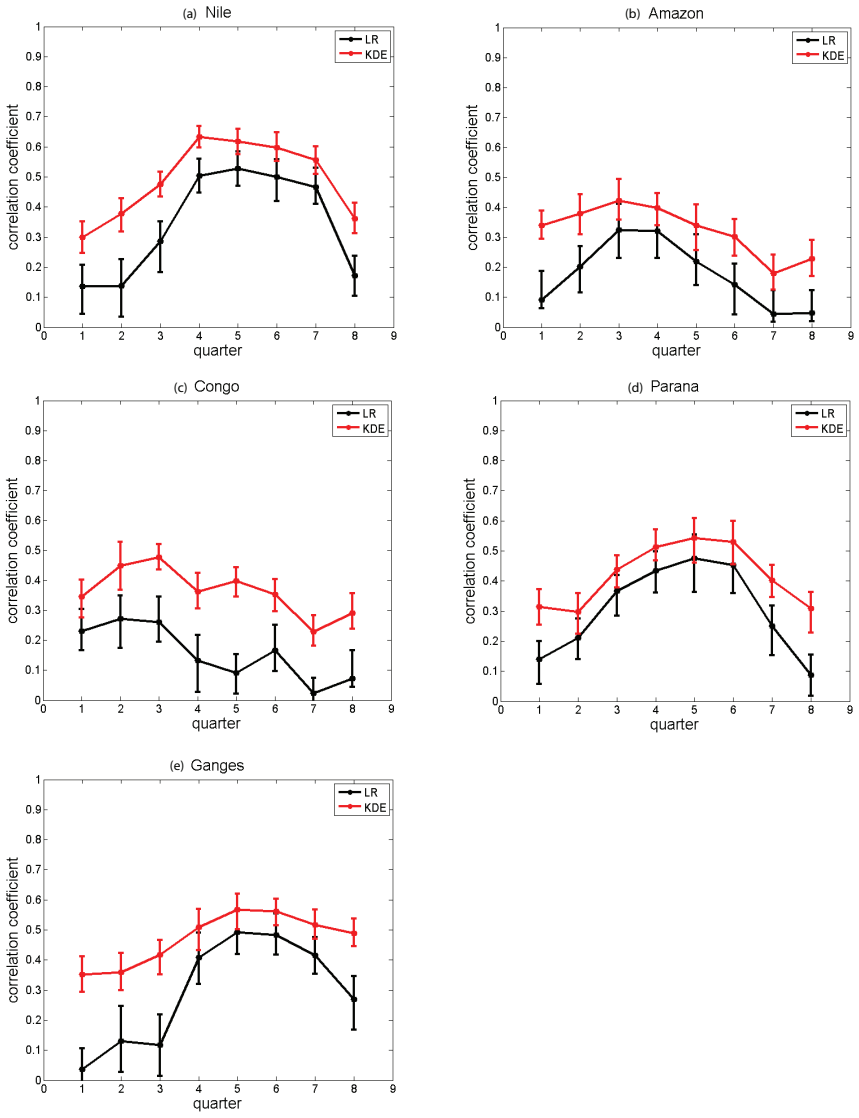

Figure 31. Nonlinear and linear CCs with their 90% confidence bounds between ENSO and annual river flows of Nile, Amazon, Congo, Paraná, and Ganges using KDE and LR approaches, respectively. The bias-corrected estimates, $(\overline{\lambda}, \overline{\rho})$, plotted as solid dots are estimated as $2(\hat{\lambda}, \hat{\rho}) - (\hat{\lambda}^*(.), \hat{\rho}^*(.))$, where $(\hat{\lambda}, \hat{\rho})$ are the original nonlinear and linear CCs between the annual flow and ENSO, respectively, considering all N observations. $(\hat{\lambda}^*(.), \hat{\rho}^*(.))$ is the mean of 100 jackknife replications of size $0.8N$ observations. The 90% confidence bounds are given by 5% and 95% quantiles of 100 jackknife replications of size $0.8N$.

4.6 Real data analysis

This dissertation assumes that the seasonal cycle for a particular year consists of 12 months starting with the month having the lowest average discharge. It also assumes that long-term flow variability due to ENSO can be captured in the annual flow, which, in turn, is defined as the integrated streamflow of the seasonal cycle. Here eight quarterly ENSO indices, i.e. three quarters just before the seasonal cycle, four quarters corresponding to the seasonal cycle, and one quarter just after the seasonal cycle, are derived from quarterly averages of mean monthly SST anomalies. The bivariate normal and kernel density between the quarterly ENSO indices and the annual flow of the Nile, Amazon, Congo, Paraná, and Ganges rivers are estimated and plotted (Figures 26-30). Linear and nonlinear CCs between the ENSO index and the annual flow of the Nile, Amazon, Congo, Paraná, and Ganges Rivers are obtained using LR and KDE, respectively (Figure 31). The bias-corrected CCs and their 90% confidence bounds are estimated using jackknifing.

The jackknife is used to estimate the bias-corrected λ and ρ and their standard errors using KDE and LR, respectively. The technique is described below for λ and is the same for ρ. In the case of real data analysis, the total number of observations (N) varies from 60 to 117. If d observations for jackknifing are left out and $\sqrt{N} < d < N$, the number of jackknife samples, given by $\binom{N}{d}$ [102], is large. So 100 samples of size $0.8N$ are used for the analysis. $\hat{\lambda}^*(.)$ and $\widehat{se}$ are the mean and standard deviation of jackknife replications leaving out $d = 0.2N$ observations. The bias is given as $\widehat{bias} = \hat{\lambda}^*(.) - \hat{\lambda}$, where $\hat{\lambda}$ is the original nonlinear CCs between the annual flow and ENSO considering all N observations. The bias-corrected estimator, $\overline{\lambda}$, is given as $\overline{\lambda} = \hat{\lambda} - \widehat{bias}$. The lower and upper bounds of 90% confidence bounds are given by 5% and 95% quantiles of 100 jackknife samples of size $0.8N$, respectively.

Table 6. Linear and nonlinear CCs between the annual flow of the Nile River and the ENSO index averaged for eight quarters. The month preceding (following) the seasonal cycle is indicated by a negative (positive) sign following a month. The bias-corrected estimates, $(\overline{\lambda}, \overline{\rho})$, are estimated as $2(\hat{\lambda}, \hat{\rho}) - (\hat{\lambda}^*(.), \hat{\rho}^*(.))$, where $(\hat{\lambda}, \hat{\rho})$ are the original nonlinear and linear CCs between the annual flow and ENSO, respectively, considering all N observations. $(\hat{\lambda}^*(.), \hat{\rho}^*(.))$ and theirs standard errors given in parentheses are the mean and standard deviation of 100 jackknife replications of size $0.8N$ observations. $\overline{\rho}$ is negative for all quarters, but the absolute values of $\overline{\rho}$ are considered. $\widehat{MSE}$ and its standard errors given in parentheses are the mean and standard deviation of MSEs estimated from 100 jackknife replications of size $0.8N$ observations.

Quarter	$\overline{\rho}$	$\widehat{MSE}_{LR}$ (10^5)	$\overline{\lambda}$	$\widehat{MSE}_{KDE}$ (10^5)
Aug.$^-$, Sep.$^-$, Oct.$^-$ (A$^-$S$^-$O$^-$)	0.135 (0.048)	2.53 (0.18)	0.300 (0.033)	2.14 (0.17)
Nov.$^-$, Dec.$^-$, Jan.$^-$ (N$^-$D$^-$J$^-$)	0.137 (0.058)	2.53 (0.20)	0.378 (0.030)	2.03 (0.16)
Feb.$^-$, Mar.$^-$, Apr.$^-$ (F$^-$M$^-$A$^-$)	0.286 (0.053)	2.34 (0.20)	0.475 (0.025)	1.82 (0.13)
May, Jun, Jul. (MJJ)	0.504 (0.035)	1.95 (0.16)	0.634 (0.023)	1.44 (0.12)
Aug., Sep., Oct. (ASO)	0.528 (0.037)	1.88 (0.15)	0.617 (0.027)	1.47 (0.12)
Nov., Dec., Jan. (NDJ)	0.501 (0.040)	1.92 (0.17)	0.597 (0.029)	1.51 (0.14)
Feb., Mar., Apr (FMA)	0.466 (0.037)	2.04 (0.14)	0.555 (0.029)	1.63 (0.11)
May$^+$, Jun.$^+$, Jul.$^+$ (M$^+$J$^+$J$^+$)	0.171 (0.042)	2.51 (0.16)	0.361 (0.030)	2.03 (0.15)

Table 7. Linear and nonlinear CCs between the annual flow of the Amazon River and the ENSO index averaged for eight quarters. The month preceding (following) the seasonal cycle is indicated by a negative (positive) sign following a month. The bias-corrected estimates, $(\overline{\lambda}, \overline{\rho})$, are estimated as $2(\hat{\lambda}, \hat{\rho}) - (\hat{\lambda}^*(.), \hat{\rho}^*(.))$, where $(\hat{\lambda}, \hat{\rho})$ are the original nonlinear and linear CCs between the annual flow and ENSO, respectively, considering all N observations. $(\hat{\lambda}^*(.), \hat{\rho}^*(.))$ and theirs standard errors given in parentheses are the mean and standard deviation of 100 jackknife replications of size $0.8N$ observations. $\overline{\rho}$ is negative from quarter 2 to quarter 6, but the absolute values of $\overline{\rho}$ are considered. $\widehat{MSE}$ and its standard errors given in parentheses are the mean and standard deviation of MSEs estimated from 100 jackknife replications of size $0.8N$ observations.

Quarter	$\overline{\rho}$	$\widehat{MSE}_{LR}$ (10^7)	$\overline{\lambda}$	$\widehat{MSE}_{KDE}$ (10^7)
Feb.$^-$, Mar.$^-$, Apr.$^-$ (F$^-$M$^-$A$^-$)	0.092 (0.063)	4.82 (0.41)	0.341 (0.027)	3.84 (0.29)
May$^-$, Jun.$^-$, Jul.$^-$ (M$^-$J$^-$J$^-$)	0.202 (0.053)	4.59 (0.45)	0.379 (0.040)	3.68 (0.33)
Aug.$^-$, Sep.$^-$, Oct.$^-$ (A$^-$S$^-$O$^-$)	0.325 (0.056)	4.29 (0.36)	0.423 (0.041)	3.49 (0.28)
Nov., Dec., Jan. (NDJ)	0.321 (0.049)	4.34 (0.38)	0.397 (0.032)	3.62 (0.30)
Feb., Mar., Apr. (FMA)	0.220 (0.052)	4.63 (0.40)	0.340 (0.039)	3.85 (0.32)
May, Jun., Jul. (MJJ)	0.141 (0.049)	4.77 (0.46)	0.302 (0.033)	3.94 (0.38)
Aug., Sep., Oct. (ASO)	0.044 (0.049)	4.85 (0.45)	0.180 (0.039)	4.16 (0.37)
Nov.$^+$, Dec.$^+$, Jan.$^+$ (N$^+$D$^+$J$^+$)	0.048 (0.053)	4.82 (0.43)	0.228 (0.036)	4.06 (0.34)

Table 8. Linear and nonlinear CCs between the annual flow of the Congo River and the ENSO index averaged for eight quarters. The month preceding (following) the seasonal cycle is indicated by a negative (positive) sign following a month. The bias-corrected estimates, $(\overline{\lambda}, \overline{\rho})$, are estimated as $2(\hat{\lambda}, \hat{\rho}) - (\hat{\lambda}^*(.), \hat{\rho}^*(.))$, where $(\hat{\lambda}, \hat{\rho})$ are the original nonlinear and linear CCs between the annual flow and ENSO, respectively, considering all N observations. $(\hat{\lambda}^*(.), \hat{\rho}^*(.))$ and theirs standard errors given in parentheses are the mean and standard deviation of 100 jackknife replications of size $0.8N$ observations. $\overline{\rho}$ is negative from quarter 1 to quarter 7, but the absolute values of $\overline{\rho}$ are considered. $\widehat{MSE}$ and its standard errors given in parentheses are the mean and standard deviation of MSEs estimated from 100 jackknife replications of size $0.8N$ observations.

Quarter	$\overline{\rho}$	$\widehat{MSE}_{LR}$ (10^7)	$\overline{\lambda}$	$\widehat{MSE}_{KDE}$ (10^7)
Nov.$^-$, Dec.$^-$, Jan.$^-$ (N$^-$D$^-$J$^-$)	0.230 (0.045)	1.74 (0.19)	0.346 (0.034)	1.39 (0.14)
Feb.$^-$, Mar.$^-$, Apr.$^-$ (F$^-$M$^-$A$^-$)	0.271 (0.054)	1.66 (0.18)	0.449 (0.045)	1.24 (0.13)
May$^-$, Jun.$^-$, Jul.$^-$ (M$^-$J$^-$J$^-$)	0.261 (0.045)	1.71 (0.19)	0.478 (0.027)	1.23 (0.13)
Aug., Sep., Oct. (ASO)	0.132 (0.054)	1.79 (0.19)	0.362 (0.036)	1.38 (0.14)
Nov., Dec., Jan. (NDJ)	0.092 (0.040)	1.83 (0.21)	0.397 (0.031)	1.37 (0.15)
Feb., Mar., Apr. (FMA)	0.167 (0.049)	1.77 (0.17)	0.352 (0.035)	1.38 (0.12)
May, Jun., Jul. (MJJ)	0.023 (0.055)	1.81 (0.19)	0.229 (0.031)	1.47 (0.14)
Aug.$^+$, Sep.$^+$, Oct.$^+$ (A$^+$S$^+$O$^+$)	0.072 (0.054)	1.79 (0.22)	0.291 (0.038)	1.42 (0.17)

Table 9. Linear and nonlinear CCs between the annual flow of the Paraná River and the ENSO index averaged for eight quarters. The month preceding (following) the seasonal cycle is indicated by a negative (positive) sign following a month. The bias-corrected estimates, $(\overline{\lambda}, \overline{\rho})$, are estimated as $2(\hat{\lambda}, \hat{\rho}) - (\hat{\lambda}^*(.), \hat{\rho}^*(.))$, where $(\hat{\lambda}, \hat{\rho})$ are the original nonlinear and linear CCs between the annual flow and ENSO, respectively, considering all N observations. $(\hat{\lambda}^*(.), \hat{\rho}^*(.))$ and theirs standard errors given in parentheses are the mean and standard deviation of 100 jackknife replications of size $0.8N$ observations. $\widehat{MSE}$ and its standard errors given in parentheses are the mean and standard deviation of MSEs estimated from 100 jackknife replications of size $0.8N$ observations.

Quarter	$\overline{\rho}$	$\widehat{MSE}_{LR}$ (10^7)	$\overline{\lambda}$	$\widehat{MSE}_{KDE}$ (10^7)
Dec.$^-$, Jan.$^-$, Feb.$^-$ (D$^-$J$^-$F$^-$)	0.141 (0.040)	1.59 (0.21)	0.315 (0.033)	1.08 (0.08)
Mar.$^-$, Apr.$^-$, May$^-$ (M$^-$A$^-$M$^-$)	0.211 (0.041)	1.54 (0.22)	0.297 (0.038)	1.08 (0.09)
Jun.$^-$, Jul.$^-$, Aug.$^-$ (J$^-$J$^-$A$^-$)	0.366 (0.043)	1.35 (0.22)	0.437 (0.032)	0.95 (0.08)
Sep., Oct., Nov. (SON)	0.435 (0.043)	1.30 (0.16)	0.513 (0.031)	0.89 (0.07)
Dec., Jan., Feb. (DJF)	0.476 (0.058)	1.25 (0.13)	0.542 (0.043)	0.85 (0.04)
Mar., Apr., May (MAM)	0.453 (0.052)	1.27 (0.13)	0.530 (0.043)	0.84 (0.06)
Jun., Jul., Aug. (JJA)	0.251 (0.055)	1.52 (0.20)	0.403 (0.033)	1.01 (0.07)
Sep.$^+$, Oct.$^+$, Nov.$^+$ (S$^+$O$^+$N$^+$)	0.087 (0.047)	1.60 (0.22)	0.309 (0.041)	1.08 (0.09)

Table 10. Linear and nonlinear CCs between the annual flow of the Ganges River and the ENSO index averaged for eight quarters. The month preceding (following) the seasonal cycle is indicated by a negative (positive) sign following a month. The bias-corrected estimates, $(\overline{\lambda}, \overline{\rho})$, are estimated as $2(\hat{\lambda}, \hat{\rho}) - (\hat{\lambda}^*(.), \hat{\rho}^*(.))$, where $(\hat{\lambda},\ \hat{\rho})$ are the original nonlinear and linear CCs between the annual flow and ENSO, respectively, considering all N observations. $(\hat{\lambda}^*(.), \hat{\rho}^*(.))$ and theirs standard errors given in parentheses are the mean and standard deviation of 100 jackknife replications of size $0.8N$ observations. $\overline{\rho}$ is negative for all quarters, but the absolute values of $\overline{\rho}$ are considered. $\widehat{MSE}$ and its standard errors given in parentheses are the mean and standard deviation of MSEs estimated from 100 jackknife replications of size $0.8N$ observations.

Quarter	$\overline{\rho}$	$\widehat{MSE}_{LR}$ (10^6)	$\overline{\lambda}$	$\widehat{MSE}_{KDE}$ (10^6)
Jul.$^-$, Aug.$^-$, Sep.$^-$ (J$^-$A$^-$S$^-$)	0.036 (0.076)	4.70 (0.55)	0.351 (0.037)	3.64 (0.46)
Oct.$^-$, Nov.$^-$, Dec.$^-$ (O$^-$N$^-$D$^-$)	0.130 (0.071)	4.83 (0.46)	0.359 (0.039)	3.76 (0.43)
Jan.$^-$, Feb.$^-$, Mar.$^-$ (J$^-$F$^-$M$^-$)	0.118 (0.063)	4.71 (0.43)	0.416 (0.033)	3.53 (0.39)
Apr., May, Jun. (AMJ)	0.408 (0.051)	4.08 (0.34)	0.509 (0.040)	3.23 (0.30)
Jul., Aug., Sep. (JAS)	0.492 (0.043)	3.56 (0.34)	0.567 (0.034)	2.82 (0.31)
Oct., Nov., Dec. (OND)	0.483 (0.040)	3.73 (0.35)	0.562 (0.028)	2.99 (0.29)
Jan., Feb., Mar. (JFM)	0.415 (0.043)	3.98 (0.36)	0.516 (0.030)	3.18 (0.31)
Apr.$^+$, May$^+$, Jun.$^+$ (A$^+$M$^+$J$^+$)	0.270 (0.053)	4.53 (0.48)	0.490 (0.029)	3.36 (0.40)

The prediction accuracies, in terms of MSEs, of annual river flows based on ENSO are also estimated and compared using LR and KDE approaches (Tables 6-10).

Table 11. Variation in the annual flow of rivers associated with ENSO. Linear and nonlinear CCs are estimated using LR and KDE, respectively. Months in a quarter are given in [ ]. The month preceding (following) the seasonal cycle is indicated by a negative (positive) sign following a month.

River	Previous Studies	Linear CC	Nonlinear CC
Nile	25% [SON] [13]	28% [ASO]	40% [MJJ]
Amazon	10% [D$^-$JF] [14]	11% [A$^-$S$^-$O$^-$]	18% [A$^-$S$^-$O$^-$]
Congo	10% [MAM] [14]	7% [F$^-$M$^-$A$^-$]	23% [M$^-$J$^-$J$^-$]
Paraná	19% [D$^-$JF] [14]	23% [DJF]	29% [DJF]
Ganges	29% [JJA] [15]	24% [JAS]	32% [JAS]

4.6.1 Description of results

Linear CCs between river flows and some quarters of the ENSO index, such as, all quarters of the ENSO index and the Nile flow, quarter 2 to quarter 6 of the ENSO index and the Amazon flow, quarter 1 to quarter 7 of the ENSO index and the Congo flow, and all quarters of the ENSO index and the Ganges flow, are negative. Since nonlinear CCs obtained from KDE, KNN, and Edgeworth do not have directionality, the absolute values of linear CCs are considered and plotted. The MI-based nonlinear dependence measure, i.e. KDE, generate higher CCs and lower MSEs as compared to linear dependence measure, i.e. LR, which shows that KDE captures more extrabasinal connection between ENSO and river flows in the tropical and subtropical regions of the world as compared to LR (Tables 11 and 6-10). The percentage variation in the annual flow of rivers associated with ENSO are calculated as the square of CCs. KDE suggests an increase of around 20-70% in the extrabasinal connection between ENSO and river flows over those suggested by LR (Figure 31 and Table 11). In the case of Nile, 90% confidence bounds of linear and nonlinear CCs are well separated for 5 quarters including quarter with the highest nonlinear CC indicating that KDE captures greater dependence between ENSO and the annual flow compared to LR (Figure 31a). KDE suggests greater dependence between the Congo flow and ENSO since 90% confidence bounds of linear and nonlinear CCs are well separated for all quarters except the first quarter (Figure 31c). In the case of Amazon, Paraná, and Ganges, 90% confidence bounds of linear and nonlinear CCs overlap for all those quarters which have higher linear CCs but for other quarters the bounds are well separated (Figures 31b, 31d, and 31e). This indicates that both KDE and LR capture nothing more than the linear dependence for some quarters based on 90% confidence bounds. However, there is an increase in the bias-corrected CCs from KDE as compared to

LR for the Amazon, Paraná, and Ganges Rivers which suggests a stronger extrabasinal connection between ENSO and the annual flow of these rivers, however with less than 90% confidence (Figures 31b, 31d, and 31e). When linear CCs are close to zero, the large difference between linear and nonlinear CCs should be interpreted with caution because of an artifact of equation 19 which scales nonlinear CCs exponentially with MI (Figures 31c and 31e).

4.6.2 Conclusion from the analysis of real data

The results with the real data reported here suggest that there exists a nonlinear extrabasinal connection between ENSO and river flows in the topics and subtropics. This dissertation also shows an appreciable increase in the variation of annual river flows linked to ENSO using nonlinear relationship measure as compared to linear measures. Hence, these results indicate additional predictability in the ENSO-streamflow extrabasinal connection when MI-based approaches are used, as compared to linear approaches used by researchers till date. The additional dependence captured by the MI-based nonlinear CCs may be useful for developing more accurate and longer streamflow models. This can, in turn, help in water resources management (e.g., reservoirs and dams for flood control, power generation, drought mitigation and preparedness for water supply).

4.7 Discussion

Streamflow series may reflect monotonic trends related to anthropogenic factors, which may include diversions, consumptions and flow regulations within the basin, in addition to possible impacts of climate change. An estimation of the likely magnitudes, as well as qualitative assessment of the evidence, of such changes may need to be performed on a case by case basis for each basin. Just as an example, three of the co-authors of this paper performed qualitative investigations for streamflows of two rivers within the United States [21]. These investigations demonstrated that meaningful studies may need to be rather time-consuming, hence such efforts are left as areas of future research for the purposes of this paper. Discussions regarding the specific datasets utilized in this paper can be found within the data sources as well as within the previous studies that have utilized these datasets [13–15, 17]. We would like to note that accounting for all the known trends, if possible, may have significant impact on the ENSO to streamflow connection. Thus, it is likely that the ENSO-streamflow extrabasinal connection is actually even higher than estimated if such trends were to be accounted for. Conversely, it is possible that some extremes are highlighted in the anthropogenic basin flow trends which tend to overemphasize the ENSO connection. On the other hand, an argument can perhaps

be made that such situations are not relevant to the point of this paper since the influence of anthropogenic or other trends will be reflected in both the linear and nonlinear measures of dependence. However, while making apriori statements may not be justified, it is likely that some of the trends will be nonlinear and that the nonlinear measures may be potentially more susceptible to the presence of outliers.

Although ENSO has a direct influence on rainfall anomalies over the tropical and subtropical regions, only a portion of the variation in the annual flow of rivers located in these regions is associated with ENSO events. This may be due to the complex relationship between rainfall and runoff, which, in turn, depends on surface hydrological and ocean-atmosphere-land interaction processes as well as noisy and potentially incomplete or corrupted data.

In recent decades, economic, population and geo-political pressures have resulted in significant changes in land-use patterns that may alter the land-atmosphere-water cycle in the tropics and subtropics. These changes in the water cycle can, in turn, impact regional precipitation, water vapor flux, and surface water flows, causing regional as well as global shifts in seasonal-to-interannual atmospheric variability. A better understanding and quantification of the relationship between ENSO and river discharges can help scientists and policy makers understand and get prepared for the changes in river discharge patterns besides attributing such changes to natural or anthropogenic drivers.

Acknowledgements

This research was partially funded by the SEED of the Laboratory Directed Research and Development Program of the Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U.S. DOE under Contract No. DE-AC05-00OR22725. The authors thank the following scientists who provided or pointed us to the data used in this research: Asst. Prof. Guiling Wang of the University of Connecticut, Dr. Earle Williams of the Massachusetts Institute of Technology (MIT), Prof. Shafiqul Islam of Tufts University, Prof. Norberto O. Garcìa of the Universidad Nacional del Litoral, Argentina, and Dr. Carlos Nobre of the Centro de Previsão de Tempo e Estudos Climáticos, Instituto Nacional de Pesquisas Espaciais, Brazil. We are thankful to Dr. Alexander Kraskov for providing us with KNN-based MI code. Helpful comments from Prof. Rafael Bras of MIT, Drs. Thomas Wilbanks, Gabriel Kuhn, Jim Nutaro, Alexander Sorokine of ORNL, and Dr. Bellie Sivakumar of the University of California, Davis, are gratefully acknowledged. The authors are thankful to an anonymous reviewer for some helpful suggestions which significantly improved the quality of the paper. In addition, the second author acknowledges the support provided by the $SensorNet^{®}$ research program at ORNL and his courtesy faculty affiliation at the University of South Florida.

Chapter 5

Spatio-temporal Variability of Daily and Weekly Precipitation Extremes in South America

Spatial and temporal variability of precipitation extremes are investigated by utilizing daily observations available at 2.5^0 gridded fields in South America for the period 1940-2004. All 65 years of data from 1940-2004 are analyzed for spatial variability. The temporal variability is investigated at each spatial grid by utilizing 25-year moving windows from 1965-2004 and visualized through plots of the slope of the regression line in addition to its quality measure (R^2). The Poisson-generalized Pareto (Poisson-GP) model, which is a peaks over threshold (POT) approach, is applied to weekly precipitation maxima residuals based on the 95%-quantile threshold, while daily data are utilized to analyze the number of consecutive daily extremes and daily extremes in a month based on the 99%-quantile threshold. Using the Poisson-GP model, we compute parameters of the GP distribution, return levels (RL) and a new measure called the *precipitation extremes volatility index* (PEVI). The PEVI measures the variability of extremes and is expressed as a ratio of return levels. From 1965-2004, the PEVI shows increasing trends in the Amazon basin except eastern parts, few parts of the Brazilian highlands, north-west Venezuela including Caracas, north Argentina, Uruguay, Rio De Janeiro, São Paulo, Asuncion, and Cayenne. Catingas, few parts of the Brazilian highlands, São Paulo and Cayenne experience increasing number of consecutive 2- and 3-days extremes from 1965-2004. The number of daily extremes, computed for each month, suggest that local extremes occur mostly from December to April with July to October being relatively quiet periods.

5.1 Introduction

Precipitation extremes can have significant impacts on human society, economics, and nature. Flooding is directly associated with precipitation extremes which can cause large number of casualties, loss of property, waterborne disease outbreaks in humans, plants and animals [103], and extensive damage to crops. An understanding of the intensity and frequency of precipitation extremes can be very useful for infrastructure development to prevent flooding and landslides, as well as for water resources and agricultural management. This may help nations and world bodies like the UN to be better prepared for future disasters caused by floods and flash floods. A better understanding of precipitation extremes can help hydrologic scientists and clima-

tologists gain enhanced understanding of precipitation processes driving the extremes and perhaps delineate possible anthropogenic or natural causes.

Previous studies investigated trends and variability of precipitation extremes in many parts of the world in the twentieth century, specifically the United States [22, 23], India [24], Southeast Asia and the South Pacific [25], Australia [23, 26], Europe [27], Caribbean [28], Italy [29], Balkans [30], Canada, Norway, Russia, China, Mexico [23], Japan [31], Sweden [32], southeastern South America [33], and the state of São Paulo, Brazil [34]. Recently the spatio-temporal variability of dependence among precipitation extremes was investigated over the entire South America for the period 1940-2004 using a new approach (suggested by Kuhn [35]) [36]. However, we are not aware of any prior investigations on spatial and temporal variability of precipitation extremes over the entire continent of South America.

Extreme value theory (EVT) has been widely used in hydrology to perform flood frequency analyses by utilizing historical records of precipitation, steamflow and other variables [104]. In recent years, EVT has been applied in multiple disciplines including hydrology [39, 44], ecology [41, 47], hurricane damage [105], temperature [106], wind speed [107], and wildfire sizes [108]. The generalized extreme value (GEV) distribution, developed by Jenkinson [37], has been traditionally utilized for modeling precipitation extremes [38–40]. This approach is also called the block maxima approach since it fits the distribution to the highest values in blocks of equal size, e.g., maximum yearly precipitation. It has some advantages, e.g., its requirements can be met by a simplified summary of data and the block maxima can be assumed to be independent random variables [41]. But the main drawback of the GEV distribution is that it does not utilize all the available information about the upper tail of the distribution, e.g., two highest extreme precipitation events may occur in the same year [41]. An alternative approach is to use peaks over threshold (POT) which was originated in hydrology and makes use of all the data available, e.g., all daily precipitation data [42]. The statistical model underlying the POT method consists of (1) Poisson process for the occurrences of extremes over a large threshold and (2) generalized Pareto (GP) distribution (with scale (σ) and shape (ξ) parameters), developed by Pickands [43], for the distribution of excesses over a large threshold. This model is also termed as Poisson-GP model. Recently, the GP distribution has been utilized for modeling threshold excesses from daily precipitation data [44,45]. This dissertation utilizes the Poisson-GP model for investigating the spatial and temporal variability of precipitation extremes at each grid point in South America.

Daily precipitation data is available in 2.5^0 gridded fields for the period 1940-2004 in South America. The Poisson-GP model assumes the data to be independent and identically distributed (IID) [39]. A long-term trend and seasonality in the data violate the assumption of identically distributed data whereas

the assumption of independent data is violated if there is temporal dependence in the data [47]. In order to check the IID assumption for the Poisson-GP model, we consider three different sets of data based on this daily data: daily data itself, weekly maxima, and weekly maxima residuals. Weekly maxima residuals are obtained by subtracting the long term mean of weekly maxima of a particular week, i.e., mean of maximum weekly precipitation across the same week for all years used in the analysis, from weekly maxima of the same week. These datasets are compared to choose the best data by examining temporal dependence through auto-correlations and seasonal trends. In order to check the quality of the Poisson-GP model, we also compare these datasets in terms of the Poisson property of the occurrences of extremes and quality of the GP distribution. The scale of the data and the need for efficient computations, which can be eventually automated, preclude choosing thresholds based on human judgment. We choose thresholds as 95%-quantile for weekly maxima and weekly maxima residuals and 99%-quantile for the daily data. The thresholds are computed at each grid, hence the extremes can be said to be local in the context of observed rainfall in the particular grid. Spatial variability is investigated for 65 years (1940-2004) and the last 40 years (1965-2004) are also studied for the temporal variability with 25-year moving window, i.e., 1965-1989, 1966-1990,. . ., 1980-2004. The temporal variability is given by the slope of a linear trend obtained by fitting a regression line to 16 values from 16 time windows from 1965-2004. We also plot R^2 obtained from fitting a regression line which provides the overall measure of the quality of fitted regression line. We investigate the spatial and temporal variability of thresholds, σ and ξ and their standard errors, 50-year and 200-year return levels (RL), and *precipitation extremes volatility index* (PEVI) which measures the variability of extremes and is defined as a ratio of RLs. This dissertation computes PEVI as the ratio of 200-year and 50-year RLs, where the latter represents a design return level, e.g., the return level used for infrastructure design, while the former represents rarer and more intense extremes. The PEVI represents a measure of surprise if the rarer extremes were to occur. The advantages of PEVI are easy interpretability, computational efficiency and effective visualization through a single parameter at each grid. The temporal variability of thresholds from 1965-2004 also gives an indication about increasing or decreasing trends in precipitation during that period. Based on daily data, the spatial and temporal variability of the number of consecutive 2-days and 3-days extremes and the spatial variations of the number of monthly extremes are investigated.

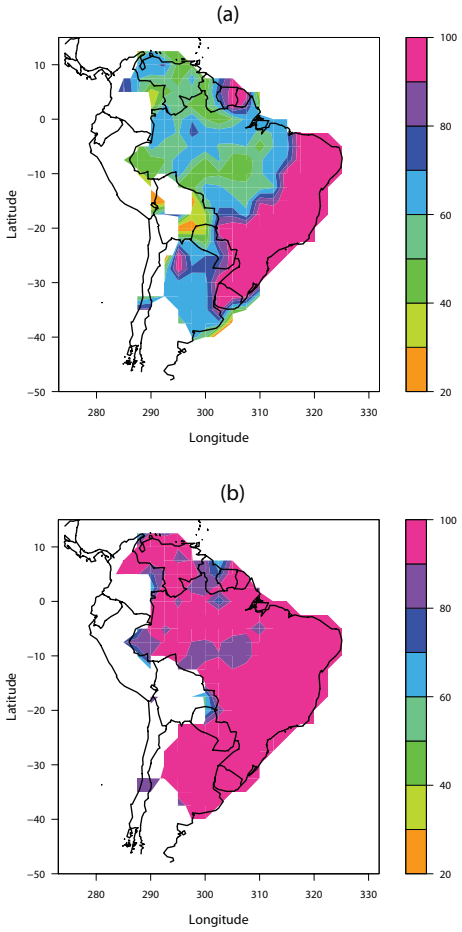

Figure 32. Percentage of total data available at each grid point: (a) Percentage of daily data available in 65 years from 1940-2004; and (b) Mean percentage of daily data available in 40 years from 1965-2004 computed using 25-year moving window from 1965-2004, i.e., 1965-1989, 1966-1990,. . ., 1980-2004. Each grid point having at least 14 years of data is considered for the analysis. This means that all grid points having more than 22% and 56% of data are used for the analysis in (a) and (b), respectively. The white regions on the map indicate either non-availability of data or insufficient data, i.e., less than 14 years of data, for the analysis.

5.2 Data and methodology

5.2.1 Data availability

The daily precipitation data used in this dissertation was published for Brazil, Venezuela, north Argentina, Paraguay, Uruguay, Suriname and French Guiana from 1940-2004 by Liebmann and Allured [109]. The data was presented in 2.5^0 gridded fields which were constructed using daily precipitation totals from 7900 stations. The daily precipitation at each point on a 2.5^0 grid was calculated by averaging daily precipitation from all stations within a radius of 1.875^0 of the point. The complete description of this data sets is given in [109]. The spatial variability is investigated for 65 years from 1940-2004 where the percentage of data points available for the analysis at each grid point is shown in Figure 32a. We analyze all those grid points having 14 or more years of data. For the spatial variability from 1940-2004, 223 grid points are analyzed since they have 14 or more years of data. We investigate temporal variability for 40 years from 1965-2004 by considering 25-year moving window, i.e., 1965-1989, 1966-1990,. . ., 1980-2004, which generates 16 values. Figure 32b shows the mean percentage of data points, i.e., mean of 16 percentages of data for 16 windows from 1965-2004, available for the analysis at each grid point. A total of 216 grid points are analyzed for the temporal variability since they have mean percentage values of 56% or more which is equivalent to 14 or more years of data out of 25 years.

5.2.2 Methodology

5.2.2.1 Poisson-GP model If $x_1, \ldots, x_n$ be a sequence of IID observations, the Poisson-GP model consists of two components: (i) the sequence of times at which exceedances occur over a large threshold u, i.e., $x_i > u$ for some i, is governed by a Poisson process; and (ii) the limiting distribution of the excesses over a large threshold u, i.e., $x_i - u$ for some i, is the GP distribution [39]. The first component implies that if threshold exceedances occur independent in time, the time intervals between threshold exceedances, also referred as inter-arrival times of threshold exceedances later in the dissertation, follow one-dimensional *homogeneous Poisson process*. By the definition of one-dimensional *homogeneous Poisson process*, the inter-arrival times of threshold exceedances are independent and exponentially distributed.

Let $x_1, \ldots, x_n$ be a sequence of IID measurements. An extreme event x is defined when it exceeds a threshold u. If $x_{(1)}, \ldots, x_{(k)}$ are the k exceedances over threshold u, then threshold excesses are defined as $y_i = x_{(i)} - u$, for $i = 1, \ldots, k$. If $y_1, \ldots, y_k$ is an independent sequence of a random variable, the distribution of these threshold excesses can be approximated by a member of the GP family [110]. The

cumulative distribution function for the GP is given by

$$F_{\sigma,\xi}(y) = \begin{cases} 1 - [1 + (\xi y/\sigma)]^{-1/\xi}, & 1 + (\xi y/\sigma) > 0, \xi \neq 0 \\ 1 - e^{-y/\sigma}, & \xi = 0 \end{cases} \tag{31}$$

where $y > 0$; $\sigma > 0$ is a scale parameter; and $-\infty < \xi < \infty$ is a shape parameter. The shape parameter is important to understand the qualitative behavior of the GP distribution. The GP distribution has an upper bound for $\xi < 0$ (also called bounded distribution) whereas it is unbounded for $\xi = 0$ (also called light-tailed distribution) and has no upper limit for $\xi > 0$ (also called heavy-tailed distribution) [41]. The parameters σ and ξ of the GP distribution are estimated by maximizing the log-likelihood function since maximum likelihood estimation assigns the highest probability to the observed data by adopting the model with the greatest likelihood out of all the models under consideration [110]. The log-likelihood function for the GP distribution defined in Equation 31 is given as

$$l_{\sigma,\xi} = \begin{cases} -k \log(\sigma) - (1 + 1/\xi) \sum_{i=1}^{k} log(c_i), & c_i > 0 \\ -k \log(\sigma) - \frac{1}{\sigma} \sum_{i=1}^{k} y_i, & \xi = 0 \end{cases} \tag{32}$$

where $c_i = (1 + \xi y_i/\sigma)$ [110]. The GP models can be easily interpreted using extreme upper quantiles or return levels. In hydrology, the return level is generally defined on an annual scale, e.g., for a return period N, N-year return level is defined as the level expected to be exceeded once in every N years, or having an exceedance probability of $1/N$ in any given year. N-year return level can be obtained by inverting Equation 31 as

$$F^{-1}_{\sigma,\xi} = RL_N = \begin{cases} u + \frac{\sigma}{\xi}[(N n_y \zeta_u)^{\xi} - 1], & \xi \neq 0 \\ u + \sigma \log(N n_y \zeta_u), & \xi = 0 \end{cases} \tag{33}$$

where u and n_y are the threshold and number of observations in a year, respectively; and $\zeta_u = k/n$ is the probability of an individual observation exceeding u [110].

The GP distribution is a limiting distribution of excesses over a large threshold, therefore the choice of threshold can be critical. If a threshold is low, it is likely to violate the asymptotic basis of the model leading to bias in estimation and extrapolation whereas a high threshold will result in small number of exceedances for model estimation leading to large estimation variance [110]. Two methods for threshold selection, which provide a reasonable approximation to the distribution of threshold excesses, are available: (a) find a threshold u_0 from the mean residual plot, which is a plot between mean of excesses and threshold u, above which the

plot is approximately linear in u, and (b) find a threshold u_0 above which the estimates of the shape parameter (ξ) and scale parameter (σ) are constant [110]. These threshold choices are based on user judgements or subjective considerations. This dissertation does not use any of the above methods for threshold selection because it is not feasible to select thresholds based on visual inspection at each and every grid point from 223 grid points in South America. Since the extreme precipitation events are assumed to be rare, the selection of a constant threshold for all spatial grid points is not recommended because this may give more number of extremes at some places or less/no extremes at other places. A spatially distributed threshold is more justifiable hydrologically because the impact of large precipitation is likely to depend on deviation from the *usual* at any given spatial location. Previous researchers chose some high quantiles, e.g., 97.5% and 95%, of the empirical distributions as thresholds [111, 112]. In this dissertation, we choose thresholds as the 99%-quantile and 95%-quantile of the daily and weekly maxima data, respectively.

5.2.2.2 Precipitation extremes volatility index (PEVI) Hydraulic structures or other civilian infrastructures are often designed to withstand extremes events of certain magnitudes. However, the infrastructures may fail if they are exposed to a rarer and more intense extreme event. A measure that compares the increase in intensity with the rarity of extreme events can be a useful indicator for vulnerability, assuming all other conditions remain the same. Assuming $T > t$, if the T-year event, which corresponds to a ($1/T$) probability of occurrence, were to be marginally higher than the t-year design event with a ($1/t$) probability, the infrastructures can be considered to be less vulnerable compared to a situation where the difference between the intensities is significant. The difference in the intensities correspond to a measure of surprise when a lower probability and more intense extreme event occurs compared to the design event. The PEVI is a new measure defined in this dissertation to quantify and visualize the anticipated surprise caused by intense extreme precipitation events. This measure is calculated here as a ratio of return levels, i.e., RL_T/RL_t, where $T > t$; RL_T and RL_t are T-year and t-year RLs, respectively. The PEVI does not contain any new information from the point of view of extreme value theory since the information contained in it can also be derived from the shape parameter (ξ) and return levels obtained from the GP distribution. The PEVI is theoretically satisfying since there is a direct relation to the shape parameter given as

$$PEVI = \frac{RL_T}{RL_t} \sim \begin{cases} (T/t)^{\xi}, & \xi > 0, \\ \log(T)/\log(t), & \xi = 0, \\ 1, & \xi < 0, \end{cases} \tag{34}$$

The PEVI takes values greater or equal to one. The engineering intuition for RL_t is that it is a *design RL* for t years corresponding to which hydraulic structures have been designed or disaster readiness or mitigation systems have been put in place. RL_T is analogous to a higher bound on the *anticipated RL* for T years which has lower probability than RL_t but may nevertheless occur in any given year. If the PEVI is unity, the higher bound on the *anticipated RL* exactly equals the *design RL* implying very less probability of more intense extremes. However, the degree of surprise, when more intense extreme occurs, increases with larger values of PEVI. In this sense, the PEVI can also be used as the safety factor for engineering design. This dissertation chooses T and t as 200 and 50 years, respectively. We also present the PEVI since it is statistically valid and can be computed relatively efficiently for each grid point. It can be more easily interpreted and visualized than the GP distribution parameters, which do not have an event-based intuitive interpretation. This provides a measure accessible not only to statisticians but also to hydrologists, climatologists, and decision-makers. The fact that the PEVI can be easily calculated and captured through a single number makes the application to high-resolution data over large geographical areas possible.

5.2.2.3 Quality of the Poisson-GP model We investigate the quality of the Poisson process by comparing the distribution of inter-arrival times of threshold exceedances with the exponential distribution. The quality of the GP distribution to threshold excesses is investigated by examining probability and quantile plots obtained by fitting the GP distribution to threshold excesses.

We compare the distribution of the inter-arrival times of threshold exceedances with the exponential distribution using the goodness-of-fit statistic D_{SP}, suggested by Michael [113], which is based on the *stabilized probability plot.* Let $t_1, \ldots, t_k$ be k inter-arrival times of exceedances over threshold given as the 95%-quantile. If $t_1 < \cdots < t_k$ is an ordered sample drawn from an exponential distribution, whose cumulative distribution function is given as $F_0(t, \lambda) = 1 - e^{-\lambda t}$ for $t \geq 0$, the stabilized plot consists of coordinates, (r_i, s_i), which can be calculated as

$$\begin{aligned} r_i &= \frac{2}{\pi} \arcsin \sqrt{\frac{1}{k}\left(i - \frac{1}{2}\right)}, \\ s_i &= \frac{2}{\pi} \arcsin \sqrt{\frac{1}{\hat{\lambda}} F_0(t_i, \hat{\lambda})}, \end{aligned}$$

where $\hat{\lambda}$ is the maximum likelihood estimator of λ under an exponential distribution. From the stabilized plot, the deviations of plotted points from a line joining (0,0) and (1,1) indicate departures from their theo-

retical values [114]. One attractive property of the stabilized plot is that the variances of plotted points are approximately equal [113]. This property motivates the definition of a goodness-of-fit statistic D_{SP} given as

$$D_{\mathrm{SP}} \quad = \quad \max_{i=1,\ldots,k} |r_i - s_i| \,,$$

which measures the maximum deviation of the plotted points from their theoretical values and removes the subjectivity in the interpretation of stabilized plots [113]. D_{SP} is analogous to and more powerful than the standard Kolmogorov-Smirnov statistic [113, 115]. D_{SP} is used here to measure the maximum deviation of the inter-arrival times of threshold exceedances from an exponential distribution. In order to test goodness-of-fit of the inter-arrival times to the exponential distribution, D_{SP} can be compared with critical values D^*_{SP}. D^*_{SP} is obtained as some sample quantile recorded from m number of samples of sample size n [114]. Coles [114] calculated D^*_{SP} as 95%-quantile of 10 000 samples of size 10, 25, and 40 data points for normal, logistic, Cauchy, and double exponential distributions. Since this dissertation analyzes 223 grid points and each grid point can have 728 to 3 380 data points, i.e., 14 to 65 years of weekly data, in the interests of computational tractability, we consider 1 000 samples for the calculation of D^*_{SP}. We generate 1 000 independent samples of sample size n and calculate $D^{\mathrm{sim}\,m}_{\mathrm{SP}}$ from the inter-arrival times of exceedances over the threshold of 95%-quantile of mth sample for each $m = 1, \ldots, 1\,000$. We choose D^*_{SP} as the 95%-quantile of $D^{\mathrm{sim}\,m}_{\mathrm{SP}}$, $m = 1, \ldots, 1\,000$. If $D_{\mathrm{SP}} \leq D^*_{\mathrm{SP}}$, the inter-arrival times of threshold exceedances are independent and exponentially distributed with 95% probability. In order to compare D_{SP} and D^*_{SP}, we define a simple measure as

$$\overline{D}_{\mathrm{SP}} \quad = \quad \frac{D_{\mathrm{SP}}}{D^*_{\mathrm{SP}}}. \tag{35}$$

If the inter-arrival times of threshold exceedances follow one-dimensional *homogeneous Poisson process*, $\overline{D}_{\mathrm{SP}} \leq 1$ with 95% probability. If $\overline{D}_{\mathrm{SP}} > 1$, we reject with 95% confidence that the inter-arrival times of threshold exceedances follow a *homogeneous Poisson process*. This dissertation uses $\overline{D}_{\mathrm{SP}}$ at all 223 grid points because it can be easily computed, plotted and visualized in space for comparisons.

The quality of the fitted GP model to threshold excesses can be assessed by probability and quantile plots. If $y_1, \ldots, y_k$ are the k excesses over a threshold u and $\hat{F}$ is an estimated GP model, the probability plot can be generated as

$$\{(i/(k+1), \hat{F}(y_i));\ i = 1, \ldots, k\},$$

where

$$\hat{F}(y) = \begin{cases} 1 - [1 + (\hat{\xi}y/\hat{\sigma})]^{-1/\hat{\xi}}, & 1 + (\hat{\xi}y/\hat{\sigma}) > 0, \hat{\xi} \neq 0 \\ 1 - e^{-y/\hat{\sigma}}, & \hat{\xi} = 0, \end{cases}$$

where $\hat{\sigma}$ and $\hat{\xi}$ are the estimated values of the scale and shape parameters, respectively. The quantile plots can be generated by plotting the points as

$$\{(\hat{F}^{-1}(i/(k+1)), y_i);\ i = 1, \ldots, k\},$$

where

$$\hat{F}^{-1}(y) = u + \frac{\hat{\sigma}}{\hat{\xi}}[y^{-\hat{\xi}} - 1].$$

Both the probability and quantile plots should consist of points lying close to the unit diagonal if the GP model is appropriate for modeling threshold excesses [110].

5.2.3 Data preparation for the validity of the Poisson-GP model

The validity of the Poisson-GP model is based on the assumption that the data should be IID. The presence of long term trends, seasonality, and temporal correlations violate the assumption of IID data [47]. Precipitation data may be temporally correlated and have long term or seasonal trends [47]. Galambos [116] investigated the effect of long term trends, seasonality, and temporal dependence in the data on the validation of extreme value theory and found that if the auto-correlation decreases as lag times increases, the asymptotic distribution of extremes is the same as that from IID samples. The detection of clustering of extremes is also important because maximum likelihood estimation technique assumes the time series of excesses over a large threshold to be independent [41]. Clustering gradually disappears as the threshold increases but there are some variables, such as temperature, which exhibit clustering even with high thresholds [110]. If there exists clusters of extremes over a high threshold, Todorovic and Zelenhasic [42] presented an ad hoc and inefficient procedure for declustering which generates a time series by choosing the highest value of each cluster.

We analyze three different sets of data generated from the daily data, such as daily, weekly maxima, and weekly maxima residuals, to choose the best data satisfying the IID assumption and improving the quality of the Poisson-GP model. Instead of presenting the results from 223 grid points, we outline the results for

two grid points representing two very different scenarios. Each of these two grid points has 65 years of daily precipitation and their locations are given in terms of $(longitude, latitude)$ as (315,-10) and (310,-25). At these two grid points, we examine time series plots for detecting seasonal trends and auto-correlation plots for detecting temporal dependence. We detect the clustering of extremes by plotting threshold excesses at these grid points. We also compare $\overline{D}_{\mathrm{SP}}$ for the quality of the Poisson process, probability and quantile plots for the quality of the GP distribution from daily, weekly maxima, and weekly maxima residuals in our quest to find the best data for the analysis. For $\overline{D}_{\mathrm{SP}}$, we analyze daily, weekly maxima, and weekly maxima residuals data for all 223 grid points in South America for three time windows, i.e., 1940-2004, 1965-1989, and 1980-2004.

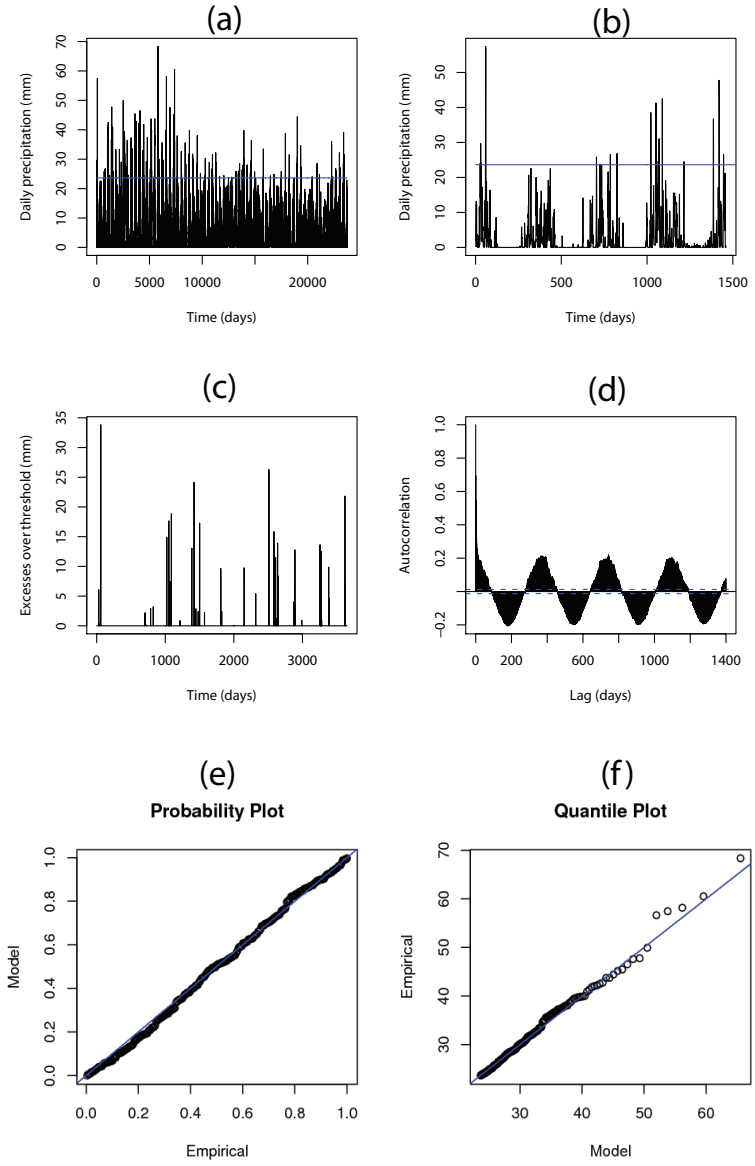

Figure 33. Grid point having $(longitude, latitude)$ as $(315, -10)$: Daily data with threshold given as 99%-quantile (shown as a horizontal line in blue in (a) and (b)). (a) Time series for 65 years; (b) Time series for 4 years; (c) Excesses over a threshold for the first 10 years; (d) Auto-correlation plot; (e) Probability plot; and (f) Quantile plot. We observe strong seasonality and temporal dependence and also some clustering of extremes. The quality of probability and quantile plots is poor.

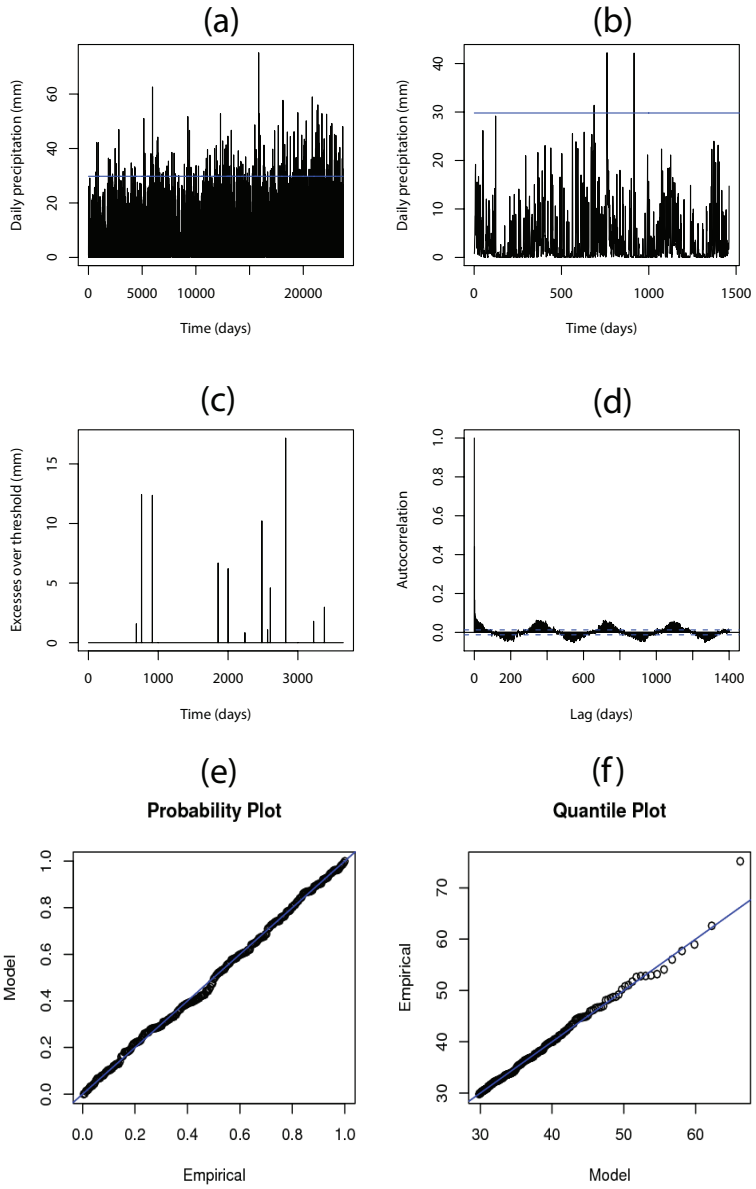

Figure 34. Grid point having $(longitude, latitude)$ as $(310, -25)$: Daily data with threshold given as 99%-quantile (shown as a horizontal line in blue in (a) and (b)). (a) Time series for 65 years; (b) Time series for 4 years; (c) Excesses over a threshold for the first 10 years; (d) Auto-correlation plot; (e) Probability plot; and (f) Quantile plot. The seasonal patterns are weak but there exists temporal dependence and clusters of extremes. The quality of probability and quantile plots is good.

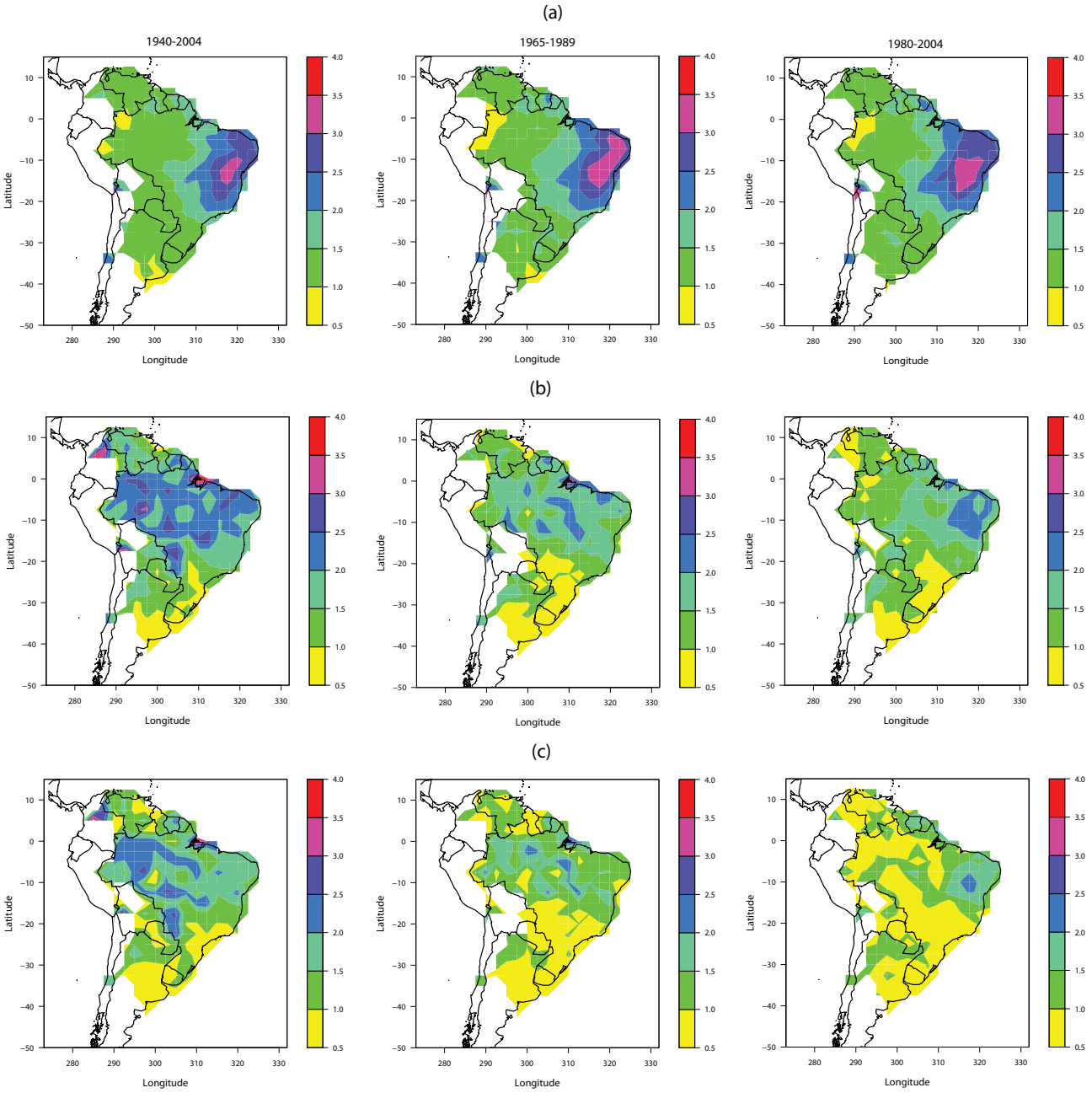

Figure 35. $\overline{D}_{\mathrm{SP}}$ for three time windows, i.e., 1940-2004, 1965-1989, and 1980-2004: (a) daily data; (b) weekly maxima; and (c) weekly maxima residuals. If $\overline{D}_{\mathrm{SP}} \leq 1$, the inter-arrival times of threshold excesses follow a *homogeneous Poisson process* with 95% probability. There is significant improvement in $\overline{D}_{\mathrm{SP}}$ from weekly maxima residuals over daily and weekly maxima data for all three time windows.

5.2.3.1 Daily We first analyze daily precipitation data to check if the IID assumption for the Poisson-GP model is satisfied. The threshold is chosen as the 99%-quantile of time series at each grid point. For 1940-2004, the time series, excesses over the threshold, and auto-correlation plots for two grid points, i.e., (315,-10) and (310,-25), are shown in Figures 33 and 34, respectively. We do not observe any long term trends at both grid points but they do show the presence of seasonality and temporal dependence. The grid point (315,-10) shows greater seasonality and temporal dependence as compared to (310,-25). A total of 88%, 85%, and 82% grid points show significant auto-correlations by visual inspection for the period 1940-2004, 1965-1989, and 1980-2004, respectively. We do observe some clusters at both grid points (Figures 33a,c and 34a,c) but do not use here the declustering method suggested by Todorovic and Zelenhasic [42] since the definition of clusters is based on subjective considerations which makes this declustering procedure inefficient for 223 grid points. $\overline{D}_{\mathrm{SP}}$ is more than one at a majority of locations in South America which means that we reject with 95% confidence that the inter-arrival times of threshold exceedances follow a *homogeneous Poisson process* at these locations(Figures 35a). The quality of probability and quantile plots obtained by fitting the GP distribution to daily data is good at (310,-25) but poor at (315,-10) (Figures 33e,f and 34e,f). Based on all the above reasons, this dissertation rules out the analysis of daily precipitation for the investigation of spatial and temporal variability of extremes using the Poisson-GP model. Therefore, we aggregate daily data into weekly maxima data in order to check if it reduces temporal dependence, resolves the clustering problem and improves $\overline{D}_{\mathrm{SP}}$ (described in the next section).

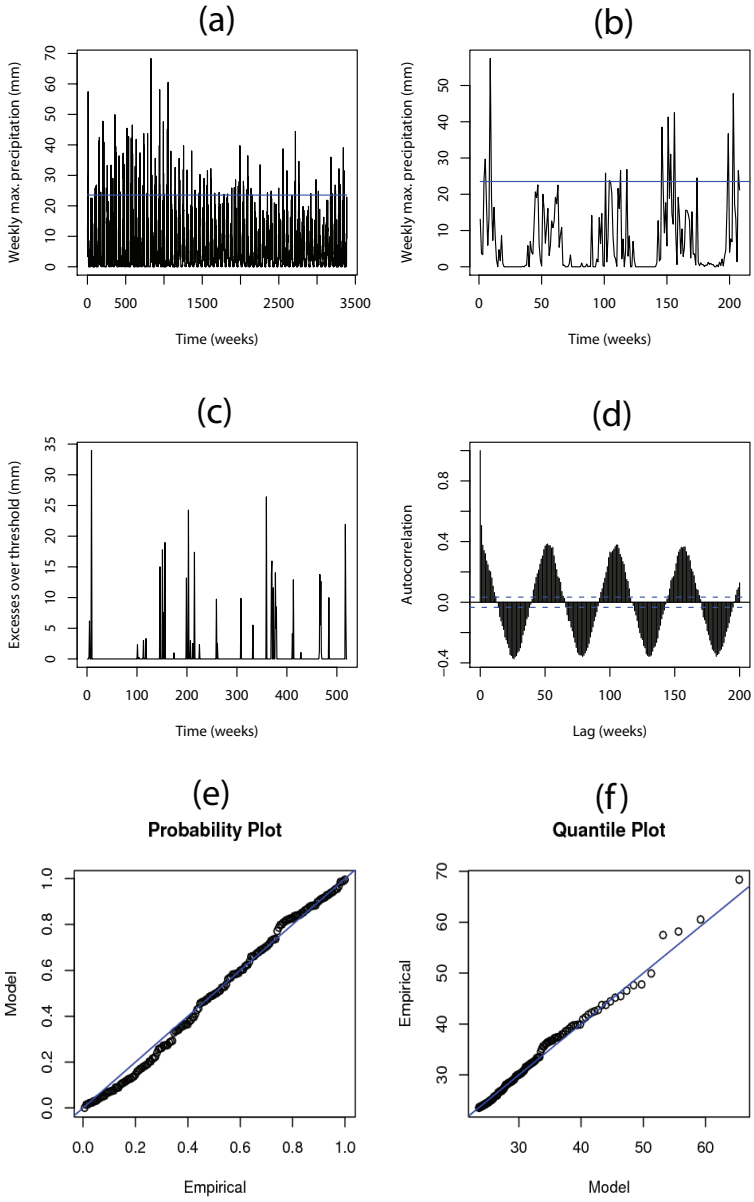

Figure 36. Grid point having $(longitude, latitude)$ as $(315, -10)$: Weekly maxima data with threshold given as 95%-quantile (shown as a horizontal line in blue in (a) and (b)). (a) Time series for 65 years; (b) Time series for 4 years; (c) Excesses over a threshold for the first 10 years; (d) Auto-correlation plot; (e) Probability plot; and (f) Quantile plot. We observe strong seasonal patterns, clusters of extremes, and temporal dependence. The quality of probability and quantile plots is bad.

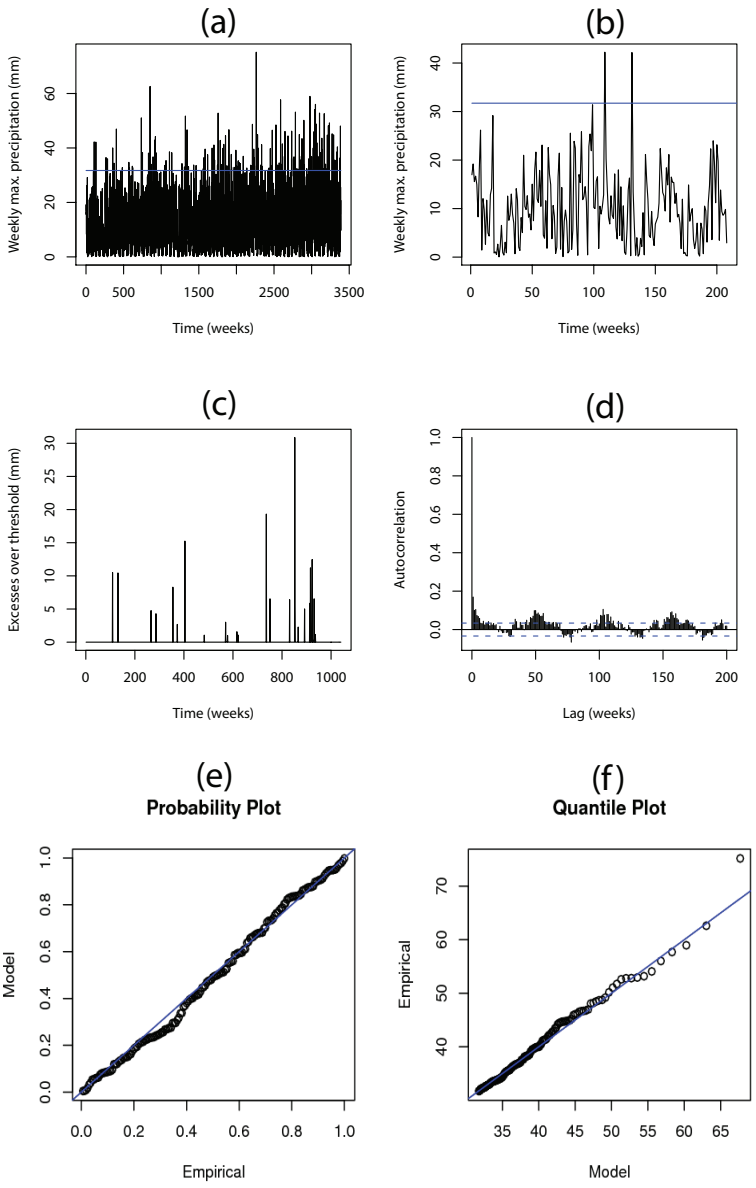

Figure 37. Grid point having $(longitude, latitude)$ as $(310, -25)$: Weekly maxima data with threshold given as 95%-quantile (shown as a horizontal line in blue in (a) and (b)). (a) Time series for 65 years; (b) Time series for 4 years; (c) Excesses over a threshold for the first 20 years; (d) Auto-correlation plot; (e) Probability plot; and (f) Quantile plot. The seasonal patterns are not evident from time series plots and there is some improvement in clustering of extremes as compared to daily data (Figure 34c). We observe some temporal dependence but it seems to be of the same order as from daily (Figure 34d). The quality of probability and quantile plots is good but not better than the plots from daily (Figures 34e,f).

5.2.3.2 Weekly maxima We generate weekly maxima precipitation time series from daily precipitation at each grid point. In this case, the threshold is chosen as the 95%-quantile of time series at each grid point. Both grid points, i.e., (315,-10) and (310,-25), do not show any long term trends (Figures 36a and 37a). At (315,-10), we do not observe significant changes in seasonality and clustering of extremes from weekly maxima as compared to daily data but weekly maxima shows greater temporal dependence than daily as shown by auto-correlation plots (Figures 33b,c,d and 36b,c,d). At (310,-25), significant improvements are observed in seasonality, clustering of extremes, and temporal dependence from weekly maxima if compared with daily (Figures 34b,c,d and 37b,c,d). We do observe some temporal dependence in weekly maxima data at (310,-25) (Figure 37d). Visual inspection of auto-correlation plots for all grid points indicates significant auto-correlations in nearly 80%, 80%, and 77% grid points for the period 1940-2004, 1965-1989, and 1980-2004, respectively. At both grid points, the quality of probability and quantile plots from weekly maxima degrades if compared with daily (Figures (33, 34, 36, 37)e,f). $\overline{D}_{\mathrm{SP}}$ shows slight improvements as compared to that from daily but it is more than one at the majority of grid points in South America (Figure 35b).

The results discussed and presented in this section provide a couple of interesting insights, which, in turn, have influenced our data analysis choices. First, minor to relatively more significant reductions, in terms seasonality or periodicity, clustering of extremes and autocorrelation, as well as improvements in terms of the $\overline{D}_{\mathrm{SP}}$ measure, are observed from the analysis of weekly maxima data compared to the corresponding daily data. This leads us to choose weekly data in this dissertation as they appear better suited to the type of extreme value analysis utilized here. The fact that daily precipitation data exhibit correlations with nearby lags in well-known and has been used, for example, in weather simulations [117]. On the other hand, the correlations are known to decay quickly and expected to be less significant at weekly time scales. While a combination of our data analysis results with known statistical insights about precipitation leads us to the choice of weekly data, we believe that analysis of daily data, potentially after creative post-processing designed to reduce the observed dependence, may yield interesting insights. However, the success of the post-processing scheme may determine our degree of belief in the results of the extreme value theory and therefore the scheme may need to get into rather involved modeling of precipitation processes. This is left as an area of future research. The second insight is that even the weekly aggregated data, while apparently more suitable than daily data, continues to retain seasonal or periodic patterns and short-term temporal dependence. This leads us to further investigate whether such patterns and dependence can be reduced from weekly data prior to extreme value analysis (described in the next section).

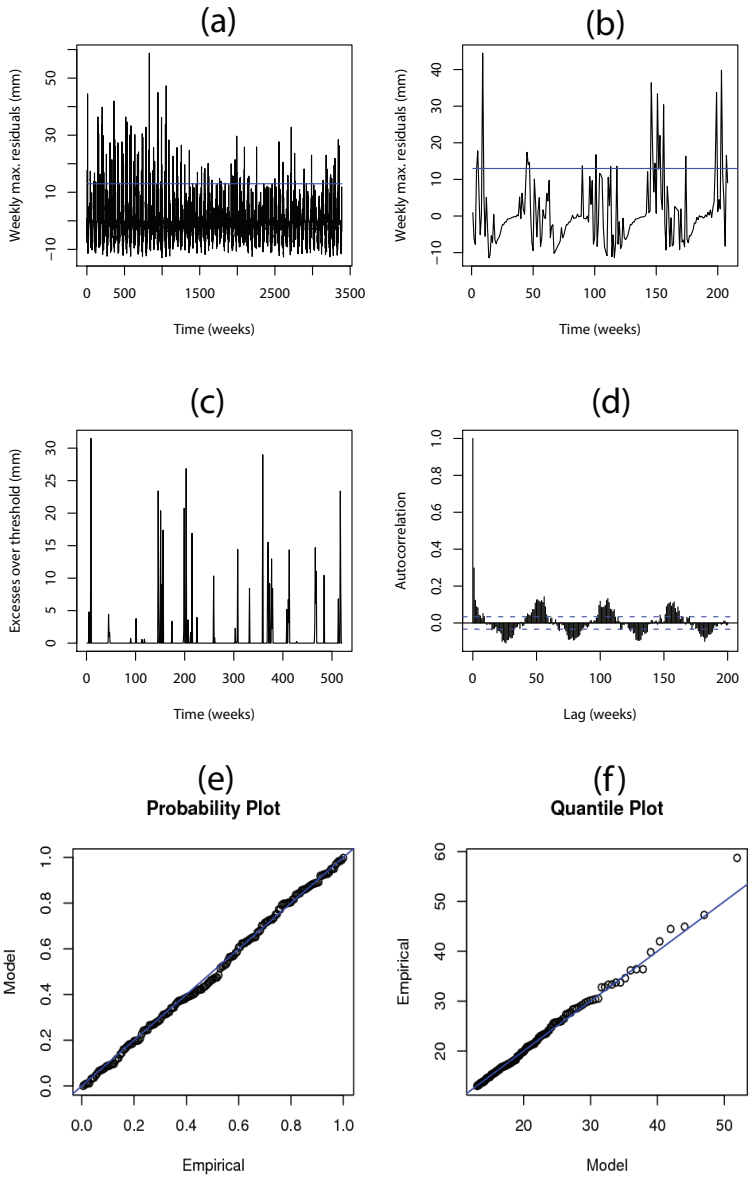

Figure 38. Grid point having $(longitude, latitude)$ as $(315, -10)$: Weekly maxima residuals data with threshold given as 95%-quantile (shown as a horizontal line in blue in (a) and (b)). (a) Time series for 65 years; (b) Time series for 4 years; (c) Excesses over a threshold for the first 10 years; (d) Auto-correlation plot; (e) Probability plot; and (f) Quantile plot. There exists strong seasonal patterns and clusters of extremes. We observe temporal dependence but it is less as compared to daily and weekly maxima (Figures 33d and 36d). The quality of probability and quantile plots is good and also better than that from daily and weekly maxima (Figures 33e,f and 36e,f).

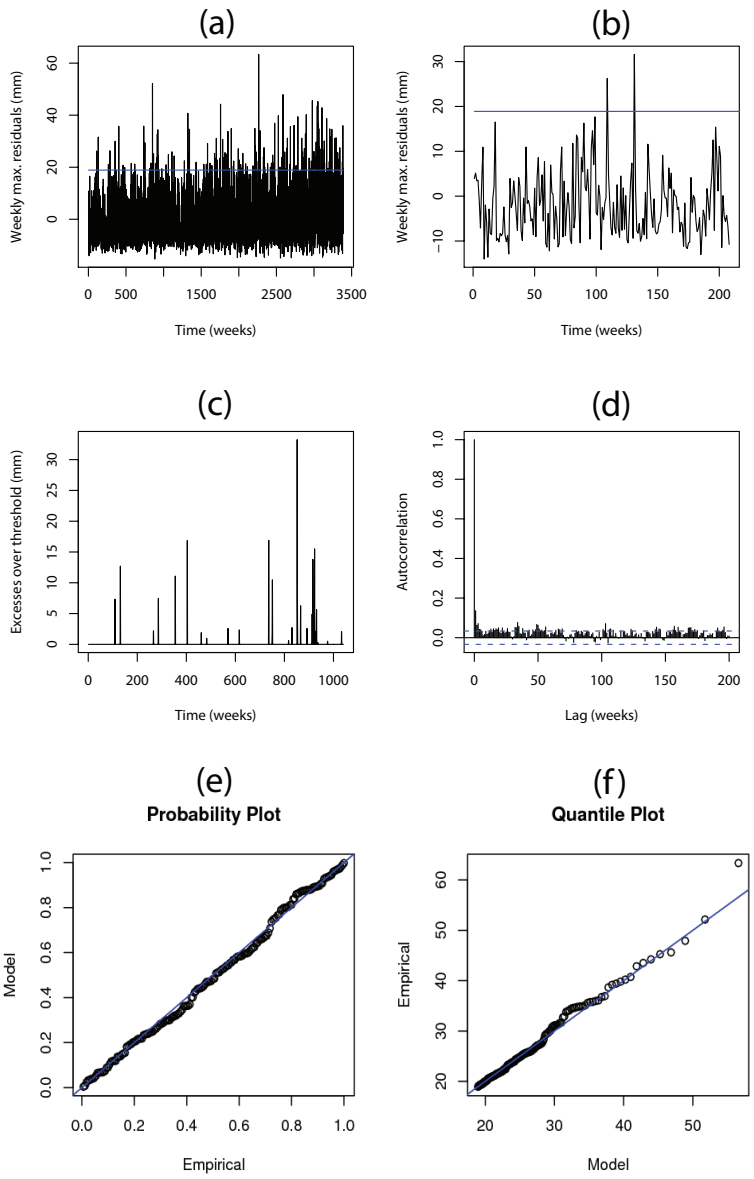

Figure 39. Grid point having $(longitude, latitude)$ as $(310, -25)$: Weekly maxima residuals data with threshold given as 95%-quantile (shown as a horizontal line in blue in (a) and (b)). (a) Time series for 65 years; (b) Time series for 4 years; (c) Excesses over a threshold for the first 20 years; (d) Auto-correlation plot; (e) Probability plot; and (f) Quantile plot. The seasonal patterns are absent. There is no improvement in clustering of extremes as compared to weekly maxima (Figure 37c). The temporal dependence disappears completely. We observe significant improvements in temporal dependence as compared to daily and weekly maxima (Figures 34d and 37d). The quality of probability and quantile plots is good.

5.2.3.3 Weekly maxima residuals We use a brute-force approach described by Gaines and Denny [47] to remove seasonality from weekly maxima to generate weekly maxima residuals. Weekly maxima residuals are obtained by subtracting the long term mean of weekly maxima of a particular week, i.e., mean of maximum weekly precipitation across the same week for all years used in the analysis, from weekly maxima of the same week. At each grid point, the threshold is chosen as the 95%-quantile of time series. Long term trends at both grid points, i.e., (315,-10) and (310,-25), are absent (Figures 38a and 39a). At (315,-10), seasonality and temporal dependence are still present for weekly maxima residuals data but it has the lowest temporal dependence as compared to daily and weekly maxima datasets (Figures (33, 36, 38)b,d). At (310,-25), clustering of extremes from weekly maxima residuals is not different from weekly maxima (Figures 37c and 39c). There is no seasonality and temporal dependence in weekly maxima residuals at (310-25) (Figures 39b,d). For weekly maxima residuals, nearly 58%, 56%, and 46% grid points show significant temporal dependence by visual inspection of auto-correlation plots for the period 1940-2004, 1965-1989, and 1980-2004, respectively, which indicates significant improvement if compared with the respective figures from daily or weekly maxima. At both grid points, probability and quantile plots consist of points lying closer to the unit diagonal indicating that the GP distribution is reasonable for modeling threshold excesses (Figures 38e,f and 39e,f). $\overline{D}_{\mathrm{SP}}$ for 1940-2004, 1965-1989, and 1980-2004 show significant improvements over daily and weekly maxima data since $\overline{D}_{\mathrm{SP}} \leq 1$ or $\overline{D}_{\mathrm{SP}}$ lies between 1 and 1.5 at the majority of places in South America for all time periods (Figures 35c). For 1965-1989 and 1980-2004, $\overline{D}_{\mathrm{SP}}$ from weekly maxima residuals is less than one in more than 50% of total grid points considered in this dissertation (Figures 35c). As we move from daily to weekly maxima residuals data for 1940-2004, $\overline{D}_{\mathrm{SP}}$ changes from greater than two to less than two for grid $(315, -10)$ whereas it changes from greater than one to less than one for $(310, -25)$ (Figures 35a and 35c). It is interesting to note that the temporal dependence go away completely if we consider weekly maxima residuals at $(310, -25)$ where $\overline{D}_{\mathrm{SP}}$ is less than one (Figures 35c and 39d). While temporal dependence persists even after considering weekly maxima residuals at $(315, -10)$ where $\overline{D}_{\mathrm{SP}}$ is between 1.5 and 2 (Figures 35c and 38d).

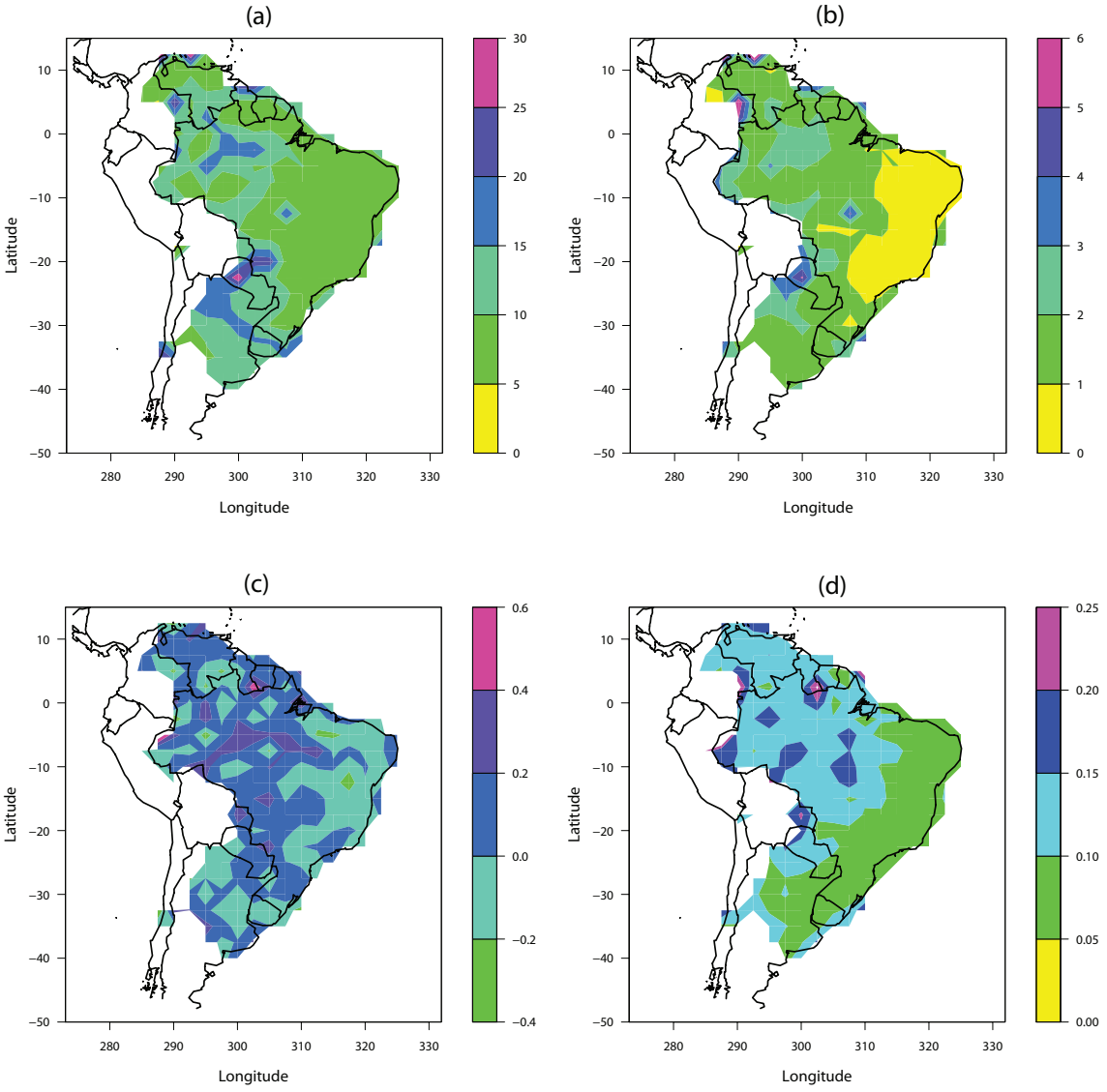

Figure 40. Scale (σ) and shape (ξ) parameters and their standard errors from weekly maxima precipitation for 1940-2004: (a) Spatial variability of σ in mm; (b) Spatial variability of standard errors of σ in mm; (c) Spatial variability of ξ; and (d) Spatial variability of standard errors of ξ.

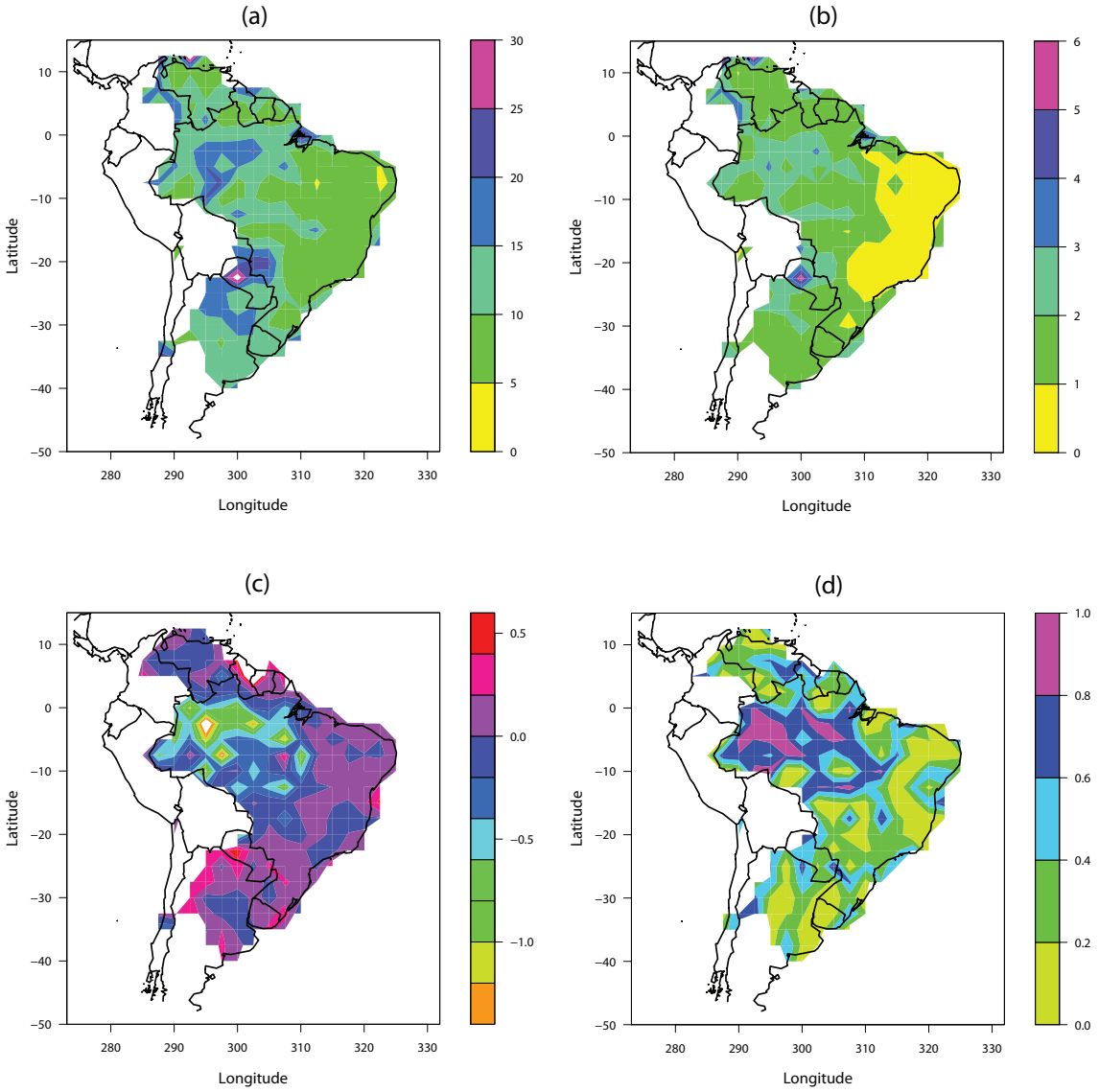

Figure 41. Scale parameter (σ) and its standard errors in mm from weekly maxima precipitation residuals: (a) Spatial variability of σ from 1940-2004; (b) Spatial variability of standard errors of σ from 1940-2004; (c) Temporal variability from 1965-2004; and (d) R^2 from linear trends shown in (c). In (c), the white region at a location given by $(longitude, latitude)$ as (295, -2.5) indicates -1.77.

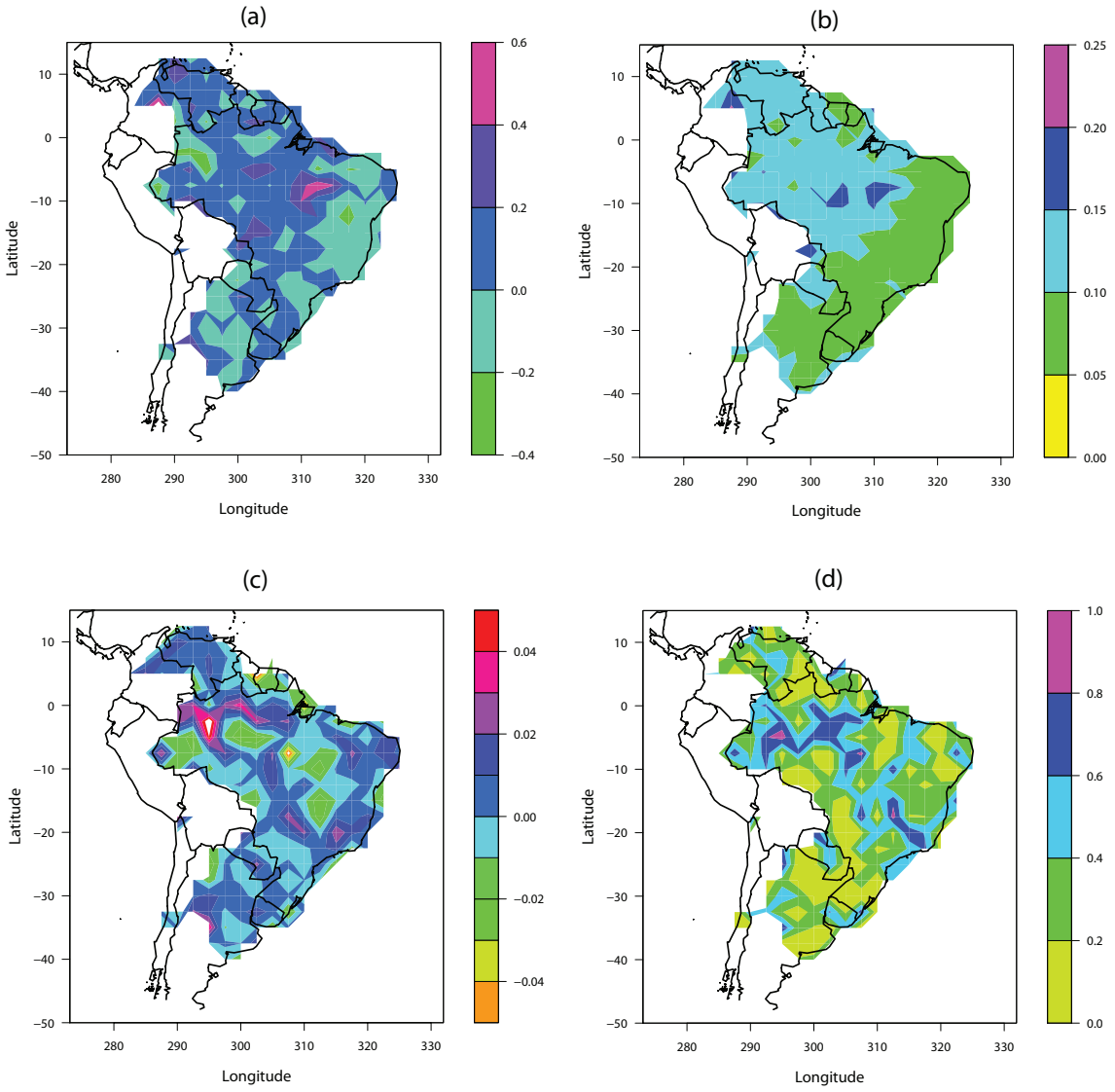

Figure 42. Shape parameter (ξ) and its standard errors from weekly maxima precipitation residuals: (a) Spatial variability of ξ from 1940-2004; (b) Spatial variability of standard errors of ξ from 1940-2004; (c) Temporal variability from 1965-2004; and (d) R^2 from linear trends shown in (c). In (c), the white region at a location given by $(longitude, latitude)$ as (295, -2.5) indicates 0.063.

The complete removal of seasonality from weekly maxima data is not possible by removing long term mean of weekly maxima from weekly maxima (Figures 38d). This process changes the order of magnitude of weekly maxima precipitation which leads to the changes in the order of magnitude of excesses over a threshold. However, both weekly maxima and weekly maxima residuals may produce most of the extremes based on their respective 95%-quantile thresholds at the same time but the excesses over their respective thresholds do not have the same magnitudes (Figures 36c, 38c, 37c and 39c). This may happen because the selection of extremes is based on 95%-quantile threshold of each time series rather than based on a fixed threshold. We fit the GP distribution to excesses over 95%-quantile threshold for both weekly maxima and weekly maxima residuals and plot the spatial variability of the scale (σ) and shape (ξ) parameters and their standard errors (Figures 40, 41a, 41b, 42a, and 42b). We observe that the spatial variability of σ and ξ from weekly maxima residuals shows the same patterns as obtained from weekly maxima in the most parts of South America. However, we observe significant improvements in the standard errors of σ and ξ from weekly maxima residuals as compared to that obtained from weekly maxima. The probability and quantile plots at two grids, i.e., (315,-10) and (310,-25), indicate better quality of the fitted GP distribution from weekly maxima residuals as compared to that from weekly maxima (Figures (36, 37, 38, 39)e,f). Since this dissertation deals with the spatial and temporal variability of extremes, we may get the same spatial and temporal patterns from both weekly maxima residuals and weekly maxima and thus, may not affect our interpretation of the results. But the analysis of weekly maxima residuals has an edge over weekly maxima because of its lower standard errors and improvements in the quality of probability and quantile plots. Weekly maxima residuals also show improvements in terms of temporal dependence and $\overline{D}_{\mathrm{SP}}$ if compared with weekly maxima. Thus, we utilize weekly maxima precipitation residuals for the analysis. Yates et al. [118] also considered weekly precipitation residuals to understand the spatial and temporal dependencies of the climate variables.

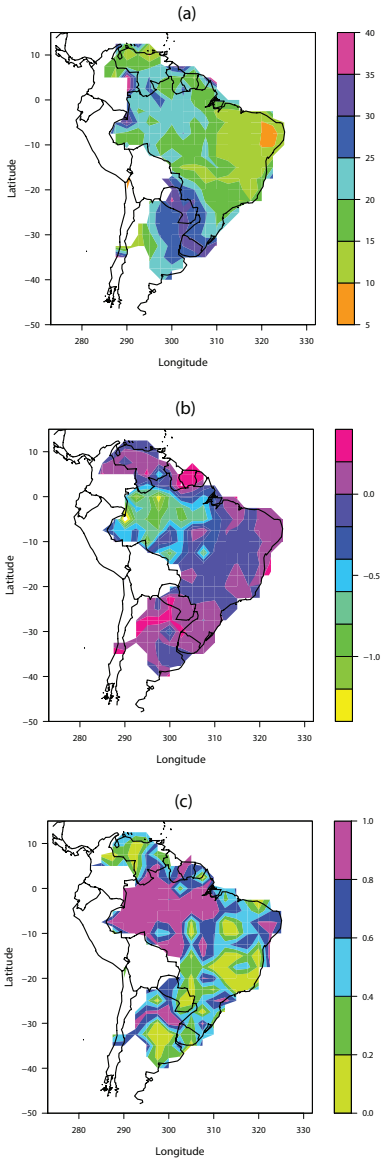

Figure 43. Threshold in mm, defined as the 95%-quantile of weekly maxima residuals at each grid point: (a) Spatial variability of threshold from 1940-2004; (b) Temporal variability at each point from 1965-2004 given as the slope of linear trend obtained by fitting a regression line to 16 threshold values computed from 25-year moving window from 1965-2004, i.e., 1965-1989, 1966-1990,. . ., 1980-2004; and (c) R^2 obtained from fitting a regression line, which provides an overall measure of the quality of linear trends shown in (b). In (b), the white region at a location given by $(longitude, latitude)$ as (290, -5) indicates -1.64.

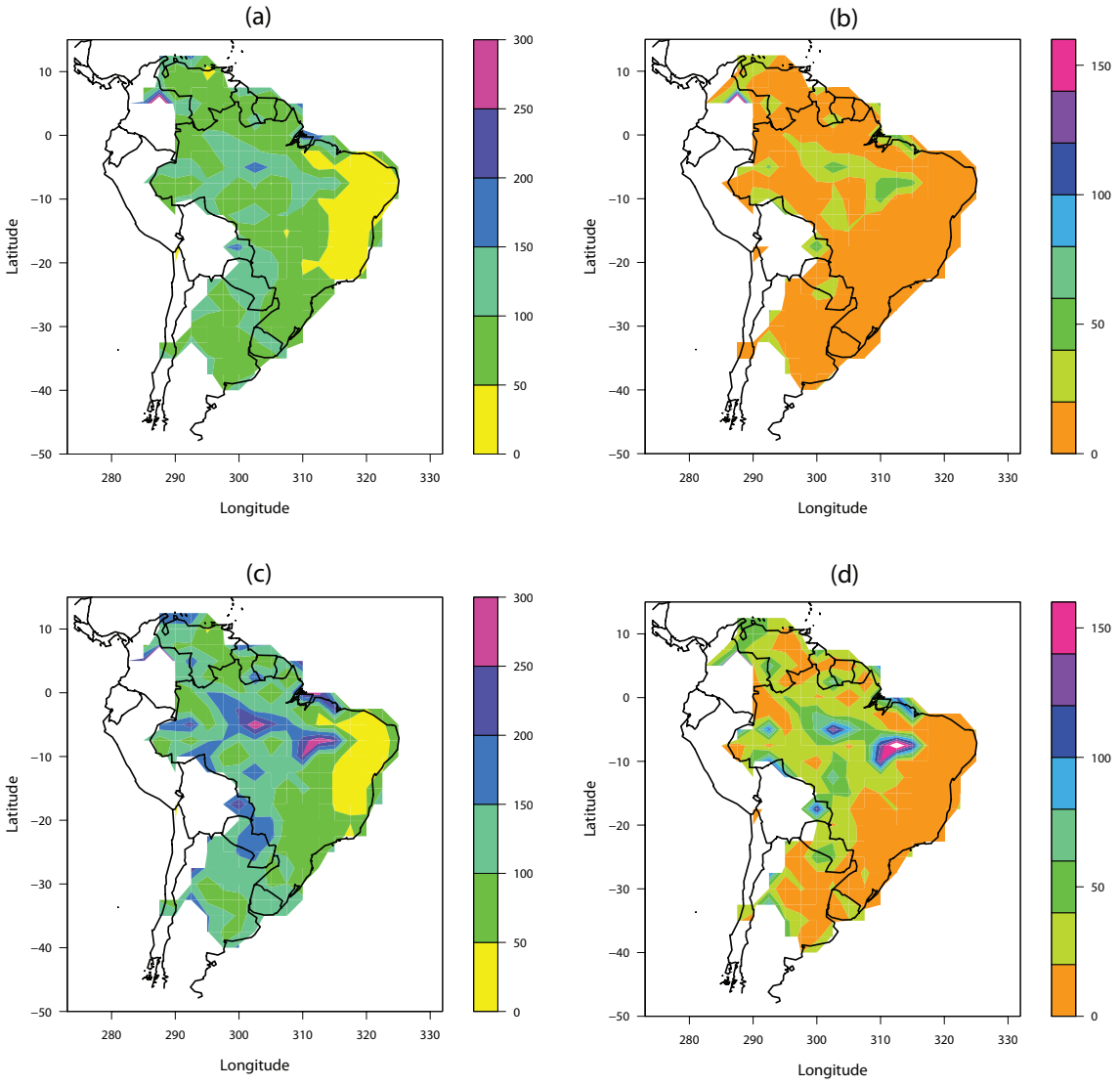

Figure 44. Spatial variability of 50-year and 200-year RLs and their standard errors in mm from weekly maxima precipitation residuals for 1940-2004: (a) 50-year RL; (b) Standard errors of 50-year RL; (c) 200-year RL; and (d) Standard errors of 200-year RL. In (d), the white region at a location given by $(longitude, latitude)$ as (312.5, -7.5) indicates 193.48 mm.

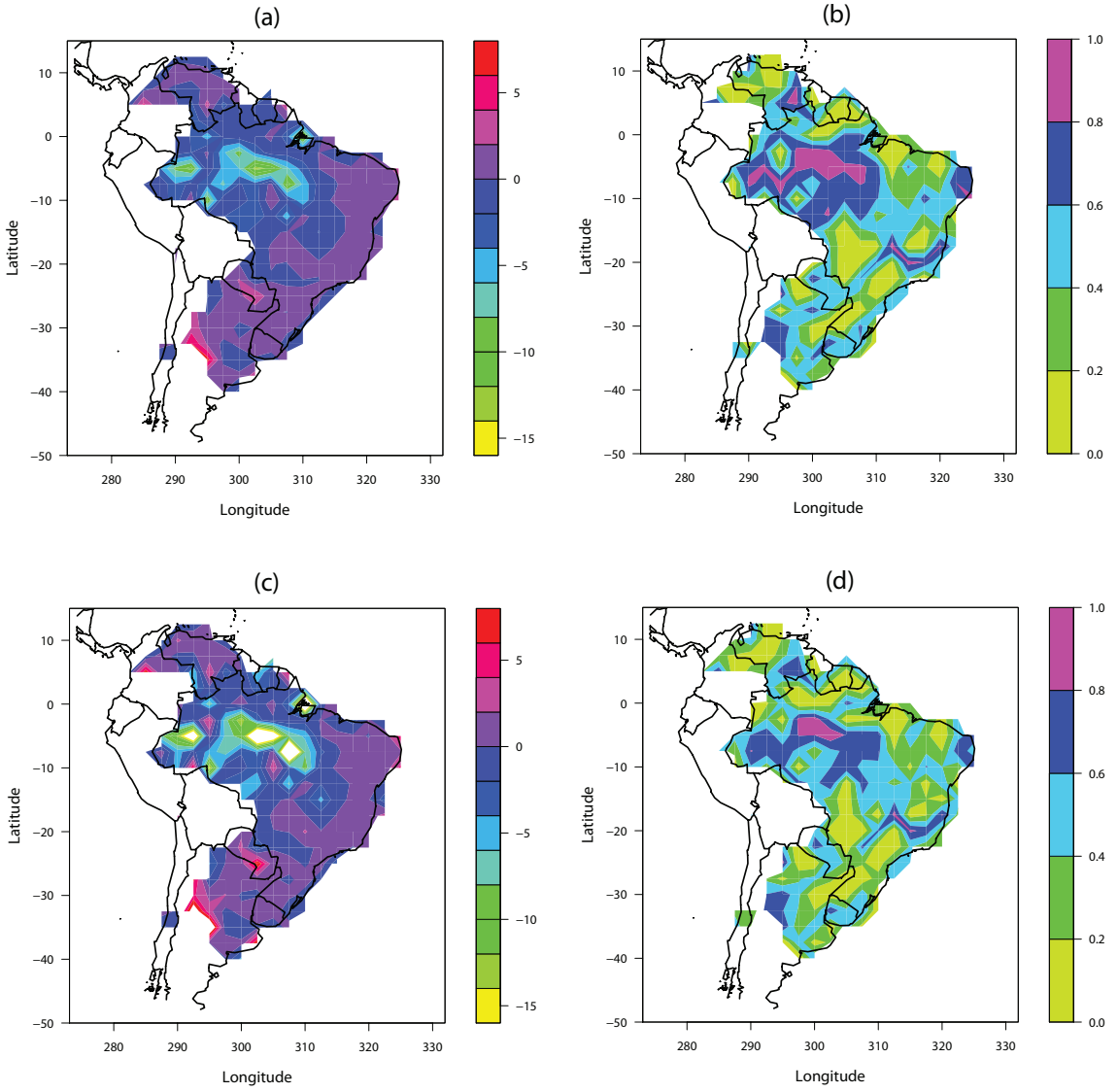

Figure 45. Temporal variability of 50-year and 200-year return levels (RL) from weekly maxima precipitation residuals for 1965-2004: (a) Temporal variability of 50-year RL from 1965-2004; (b) R^2 from linear trends shown in (a); (c) Temporal variability of 200-year RL from 1965-2004; and (d) R^2 from linear trends shown in (c). In (c), the white regions at four locations given by $(longitude, latitude)$ as (292.5, -5), (302.5, -5), (305, -5) and (307.5, -7.5) indicate -22.07, -26.39, -21.12, and -34.47, respectively.

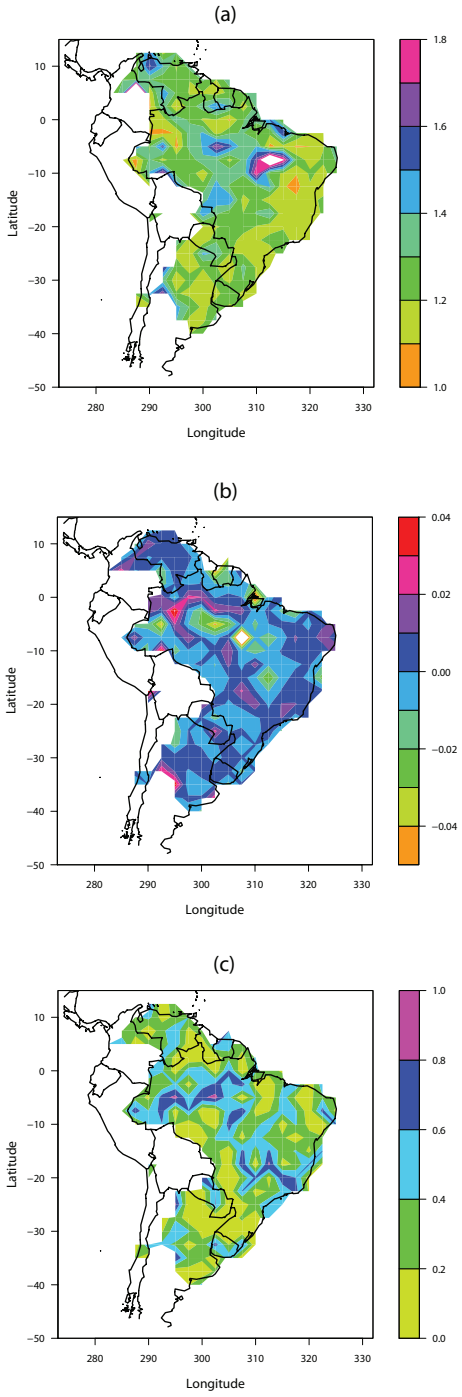

Figure 46. Precipitation extremes volatility index (PEVI), defined as the ratio of 200-year and 50-year RLs, from weekly maxima precipitation residuals: (a) Spatial variability for 1940-2004; (b) Temporal variability from 1965-2004; and (c) R^2 from linear trends shown in (b). In (a), the white regions at two locations given by $(longitude, latitude)$ as (312.5, -7.5) and (315, -7.5) indicate 2.22 and 1.82, respectively. In (b), the white region at a location given by $(longitude, latitude)$ as (307.5, -7.5) indicates 0.042.

5.3 Results and discussions

We analyze weekly maxima residuals to investigate the spatial and temporal variability of threshold, 50-year RL, 200-year RL, and PEVI (Figures 43, 44, 45, and 46). Increasing or decreasing trends in precipitation from 1965-2004 can be evaluated from the temporal variability of thresholds. At each grid point, the threshold is chosen as the 95%-quantile of time series. Spatial variability is investigated for 65 years (1940-2004) and the last 40 years (1965-2004) are analyzed for the temporal variability, which is given as the slope of linear trend obtained by fitting a regression line to 16 values computed from 25-year moving window from 1965-2004, i.e., 1965-1989, 1966-1990,. . ., 1980-2004. Using the GP distribution, the spatial and temporal variations in σ and ξ and their standard errors are evaluated and shown in Figures 41 and 42. σ ranges from 5-15 mm at major parts of South America except some parts of the Amazon basin, north Argentina, and Paraguay where it is more than 15 mm whereas the standard error of σ varies from 0-3 mm in the whole South America (Figure 41a and 41b). From 1965-2004, σ increases in eastern Brazil including Rio De Janeiro and major parts of the Brazilian Highlands, Uruguay, Paraguay, some parts of north Argentina, south Venezuela, French Guiana and Suriname whereas decreasing trends in σ are observed in the Amazon basin, Venezuela, the Mato Grasso Plateau, Catingas, São Paulo, and Buenos Aires (Figure 41c). The shape parameter (ξ) is mostly greater than zero in the whole South America except eastern Brazil and north Argentina (Figure 42a). ξ ranges from 0.4-0.6 and 0.2-0.4 in the Catingas and Mato Grasso Plateau, respectively. The standard error of ξ varies from 0.05-0.15 in the whole South America (Figure 42b). From 1965-2004, the temporal variations in ξ indicate increasing trends in Venezuela, eastern Brazil including Rio De Janeiro and São Paulo, and major parts of the Amazon basin, the Brazilian Highlands, Uruguay, Paraguay, and north Argentina including Buenos Aires (Figure 42c).

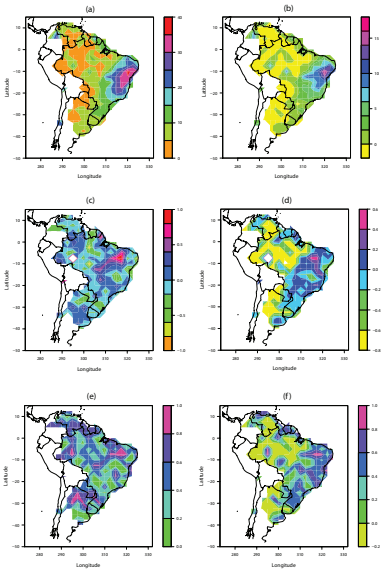

Figure 47. (Please look at the last figure for an enlarged one) Percentage of the number of consecutive 2- and 3-days extremes out of the total number of extremes based on daily precipitation for 1940-2004. Threshold is chosen as the 99%-quantile of daily time series. (a) Spatial variability of consecutive 2-days extremes from 1940-2004; (b) Spatial variability of consecutive 3-days extremes from 1940-2004, where the yellow regions showing values between 0 and -2 do not indicate any values but represents regions where the number of consecutive 3-days extremes is zero; (c) Temporal variability of consecutive 2-days extremes from 1965-2004; (d) Temporal variability of consecutive 3-days extremes from 1965-2004, where the yellow regions showing values between -0.6 and -0.8 do not indicate any values but represents regions where the number of consecutive 3-days extremes is zero; (e) R^2 from linear trends shown in (c); and (f) R^2 from linear trends shown in (d), where the yellow regions that lies between 0 and -0.2 do not indicate R^2 values but represents grids where the number of consecutive 3-days extremes is zero. In (c), the white region at a location given by $(longitude, latitude)$ as (295, -7.5) indicates 4.66. In (d), the white regions at two locations given by $(longitude, latitude)$ as (295, -7.5) and (302.5, -10) indicate 4.71 and -0.97, respectively.

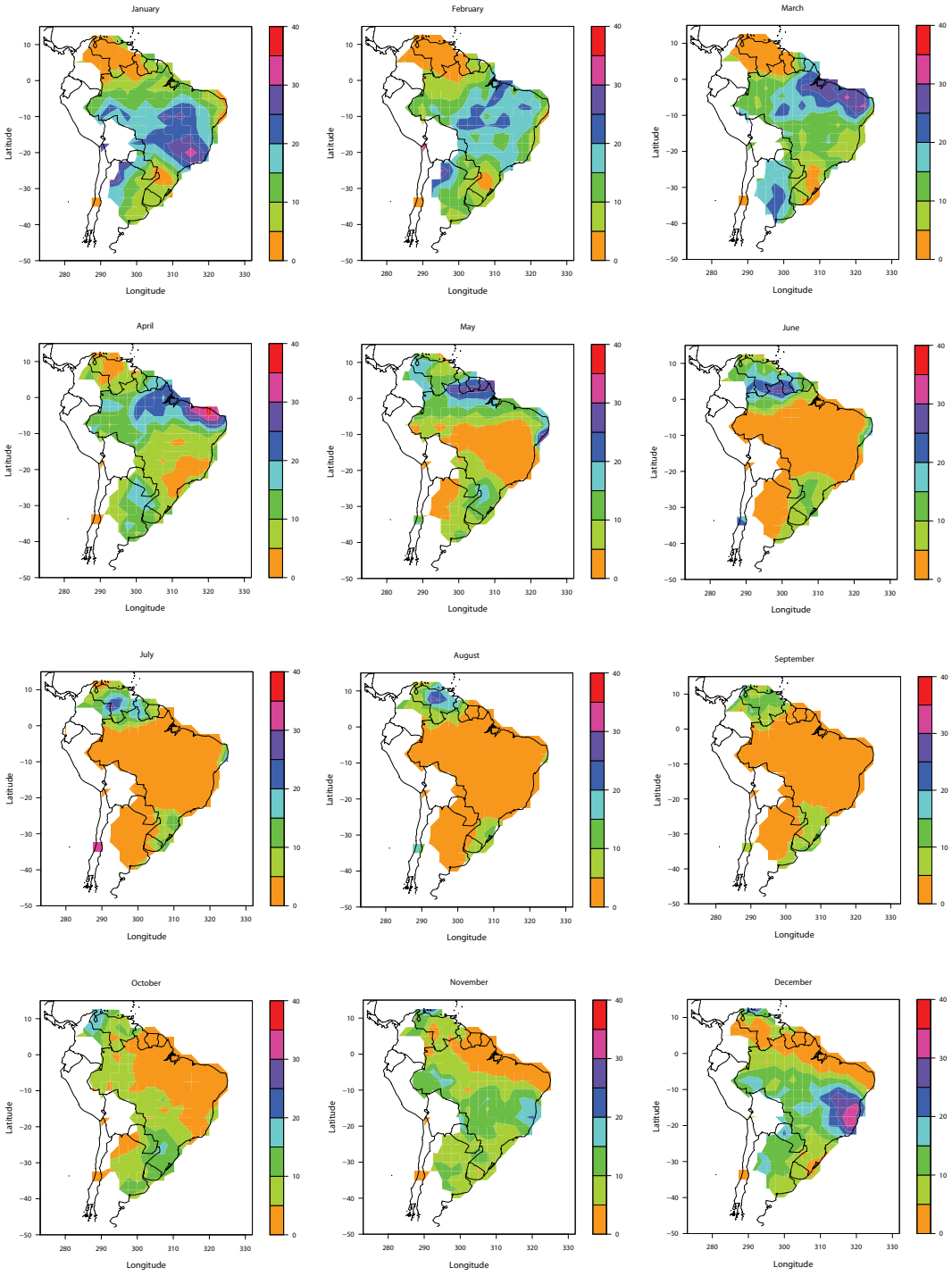

Figure 48. Percentage of the number of monthly extremes out of the total number of extremes based on daily precipitation for the period 1940-2004. Threshold is chosen as the 99%-quantile of daily time series. Extremes mostly occur from December to April with January receiving the highest number of extremes. The period from July to October is relatively quieter with respect to extremes.

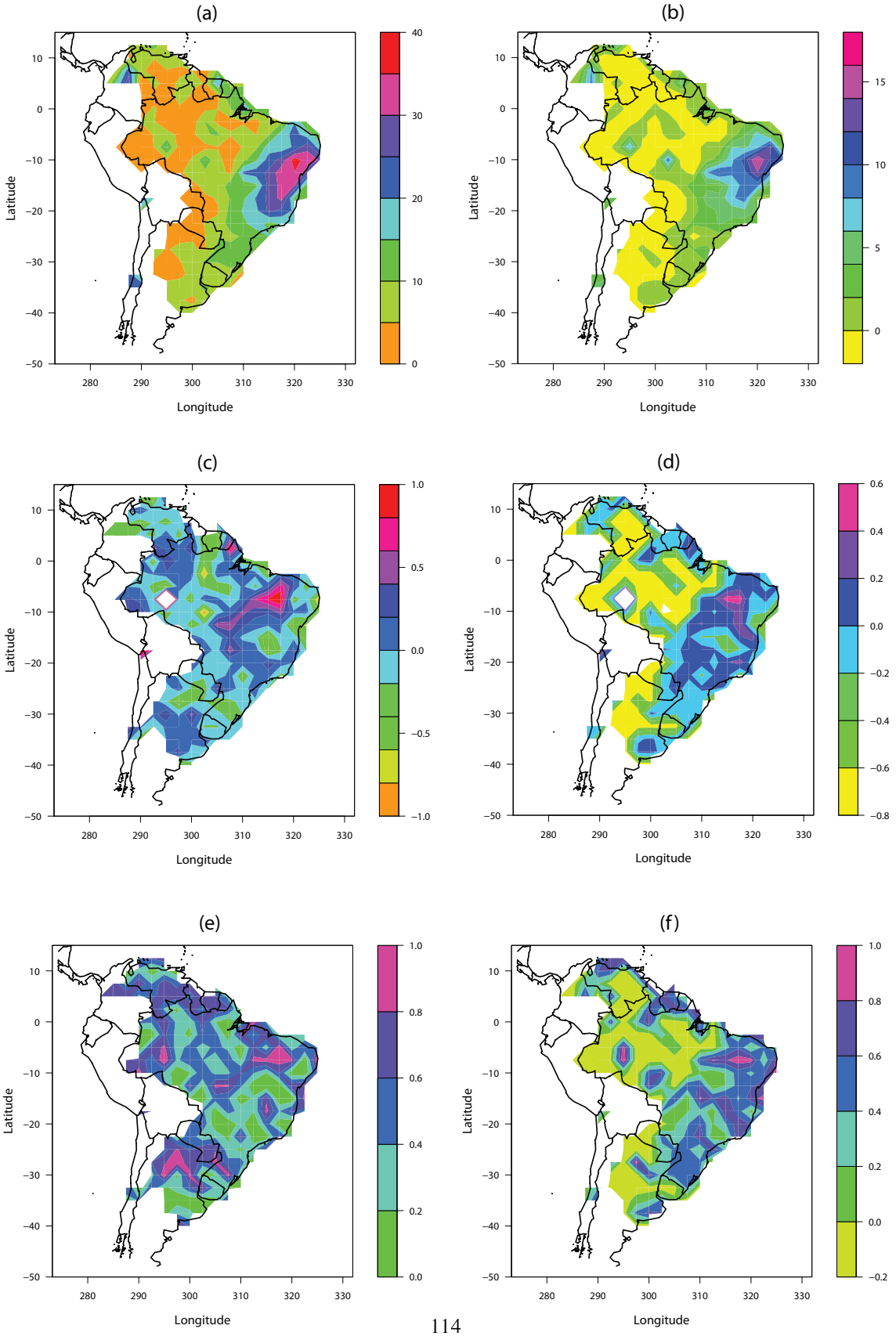

Figure 49. Same as Figure 47.

The daily data is analyzed to investigate the spatial and temporal variability of consecutive 2- and 3-days extremes (Figure 47). In this case, the threshold is chosen as the 99%-quantile of daily time series. The consecutive 2- and 3-days are defined in terms of the percentage of the number of extremes occurring consecutively for 2- and 3-days out of the total number of extremes. We also investigate the spatial variations of monthly extremes which is defined as the percentage of the number of extremes occurring in a particular month out of the total number of extremes (Figure 48).

Individual nations need to make policy decisions about water resources, agricultural planning, infrastructure management and disaster readiness or mitigation strategies. Thus, we present our results by countries in *3.1-3.6*. An investigation of precipitation extremes in conjunction with topography and vegetation, which is presented in *3.7*, can lead to enhanced hydrological and climatological insights.

5.3.1 Brazil

In the Amazon basin, threshold is larger than the other parts of Brazil but shows a decreasing trend from 1965-2004 (Figure 43). In the eastern parts of the Amazon basin, 50-year and 200-year RLs and their standard errors are higher than the other parts of South America but these RLs decrease more sharply as compared to the other parts of South America from 1965-2004 (Figures 44 and 45). Both 50-year and 200-year RLs show decreasing trends from 1965-2004 in the whole basin (Figure 45). The PEVI is higher in some eastern parts of the basin but it decreases sharply from 1965-2004 in those parts (Figure 46). We observe increasing PEVI trends from 1965-2004 in the major parts of the basin including north-west (NW) where it shows sharply increasing trends. The percentage of the number of consecutive 2-days extremes is less than 10% whereas the major parts of the basin have zero number of consecutive 3-days extremes (Figures 47a and 47b). From 1965-2004, the percentage of consecutive 2-days extremes increases only in the western parts of the basin whereas no trends in the percentage of consecutive 3-days extremes are observed because of zero number of consecutive 3-days extremes in the basin (Figures 47c and 47d).

In Catingas and the Mato Grasso Plateau, thresholds are lower and show decreasing trends from 1965-2004 (Figure 43). Catingas has the highest 50-year RL, 200-year RL, and PEVI but their trends indicate downward behavior from 1965-2004 (Figures 44, 45, and 46). In the Mato Grasso Plateau, the PEVI is higher relative to the major parts of South America and lies between 1.4 and 1.5 but it shows a decreasing trend from 1965-2004 (Figure 46). The number of consecutive 2- and 3-days extremes are less than 10% and 2% of the total extremes, respectively, in both Catingas and the Mato Grasso Plateau (Figures 47a and 47b). From 1965-2004, the percentage of consecutive 2-days extremes shows increasing trends in Catingas

and some parts of the Mato Grasso Plateau whereas the percentage of consecutive 3-days extremes indicates increasing and deceasing trends in Catingas and the Mato Grasso Plateau, respectively (Figures 47c and 47d).

In the Brazilian Highlands, threshold is low but it shows an increasing trend in the southern parts (Figure43). Thresholds are low in east Brazil except south eastern Brazil where they are much higher relative to the other parts of South America. From 1965-2004, we observe increasing trends in threshold along eastern coastal regions of Brazil including Rio De Janeiro but thresholds show decreasing trends in Brasilia, São Paulo and their surrounding regions. 50-year RL, 200-year RL, and PEVI are low in the Brazilian Highlands and east Brazil but they show increasing trends in the major parts of the Brazilian Highlands and east Brazil including Rio De Janeiro and São Paulo (Figures 44, 45, and 46). In Brasilia, decreasing trends in PEVI are observed from 1965-2004 (Figure 46). 20-35% and 6-16% of the total extremes occur consecutively for 2 and 3 days, respectively, in the Brazilian highlands and north-east (NE) Brazil (Figures 47a and 47b). In Rio De Janeiro and São Paulo, 15-20% of the total extremes occur for 2 days consecutively (Figure 47a). From 1965-2004, the number of consecutive 2- and 3-days extremes show increasing trends in NE Brazil, few parts of the Brazilian Highlands and east Brazil including São Paulo but they decrease in Rio De Janeiro (Figures 47c and 47d). In Brasilia, the number of consecutive 2-days and 3-days extremes vary 20-25% and 6-8%, respectively, and they show decreasing trends from 1965-2004 (Figure 47).

The Amazon basin experiences most of the extremes from January to April with March being the month most prone to extremes while it receives most of the rainfall from December to May (Figure 48). The wetter months in NE Brazil and Catingas are from December to May but they receive most of the extremes from January to April with March receiving the highest number of extremes. The Mato Grosso Plateau experiences most extremes from December to February with January receiving the highest number of extremes. The Brazilian highlands and south-east (SE) Brazil receives most of the rainfall from November to April but the highest number of extremes are observed in the summer months, i.e., December to February, with January being the most critical with respect to the number of extremes.

5.3.2 North Argentina

Some parts of NE Argentina including Buenos Aires have higher thresholds relative to the other parts of South America (Figure 43). From 1965-2004, increasing trends in threshold are observed in the major parts of north Argentina excluding Buenos Aires. 50-year and 200-yr RLs do not show much variations and their trends from 1965-2004 show increasing behavior in the major parts excluding Buenos Aires (Figures 44 and 45). The PEVI ranges from 1.1-1.2 in the major parts but it also lies between 1.2 and 1.3 in some parts

including Buenos Aires (Figure 46). Only some parts of north Argentina excluding Buenos Aires show increasing trends in PEVI from 1965-2004. Less than 10% of the total extremes occurs consecutively for 2 days (Figure 47a). The number of consecutive 3 days extremes is zero everywhere except in Buenos Aires and its surrounding regions where it is less than 2% (Figure 47b). Increasing trends in consecutive 2-days extremes are observed in the major parts from 1965-2004 (Figure 47c). For consecutive 3-days extremes, most of the areas experiences either no trends because of zero consecutive 3-days extremes or decreasing trends from 1965-2004 (Figure 47d). From 1965-2004, Buenos Aires receives increasing and decreasing number of consecutive 2-days and 3-days extremes, respectively. Argentina receives most of the rainfall in the summer months, i.e., December to February, but it experiences most of the extremes in late summer and autumn, i.e., February to April, with March receiving the highest number of extremes (Figure 48).

5.3.3 Venezuela

In Venezuela, thresholds are low but they show increasing trends from 1965-2004 everywhere except in some southern parts (Figure 43). We do not observe much variations in 50-year and 200-year RLs but these RLs show increasing trends from 1965-2004 only in the northern parts including Caracas (Figures 44 and 45). The PEVI is between 1.2 and 1.3 everywhere except in NW including Caracas where it is high and ranges from 1.3-1.6 (Figure 46). Increasing trends in PEVI are observed from 1965-2004 in the major parts including Caracas. Some northern parts including Caracas receive 5-10% of the total extremes for 2 days consecutively (Figure 47a). Venezuela does not experience extremes occurring for 3 days consecutively (Figure 47b). From 1965-2004, the number of consecutive 2-days extremes shows increasing trends only in south Venezuela whereas no trends are observed in consecutive 3-days extremes in the major parts since these parts do not receive consecutive 3-days extremes (Figures 47c and 47d). Decreasing trends in consecutive 3-days extremes are observed from 1965-2004 only in east Venezuela. The main rainy season in Venezuela is from May to November and it receives the most number of extremes from June to August with June being the most critical with respect to the number of extremes (Figure 48).

5.3.4 Uruguay

Uruguay has high thresholds in South America and their trends from 1965-2004 indicate increasing levels in the most parts except Montevideo and its surrounding regions (Figure 43). We do not observe much variations in 50-year and 200-year RLs and their trends show increasing behavior from 1965-2004 everywhere except in Montevideo and its surrounding areas (Figures 44 and 45). The PEVI ranges from 1.2-1.3 everywhere

and it shows increasing trends from 1965-2004 in the whole country except Montevideo and its surrounding regions (Figure 46). 5-10% of the total extremes occur consecutively for 2 days whereas less than 2% of the total extremes occur for 3 days consecutively (Figures 47a and 47b). From 1965-2004, decreasing trends in the number of both consecutive 2- and 3-days extremes are observed (Figures 47c and 47d). Uruguay receives most of the rainfall in the autumn months, i.e., March to May, but the highest number of extremes are observed in April and October (Figure 48).

5.3.5 Paraguay

In Paraguay, thresholds are high and they show increasing trends from 1965-2004 in most parts including Asuncion (Figure 43). 50-year RLs do not vary much whereas some variations are observed in 200-year RLs (Figure 44). Increasing trends in both 50-year and 200-year RLs are observed from 1965-2004 but these trends increase more rapidly in Asuncion and its surrounding areas as compared to the other parts (Figure 45). The PEVI varies from 1.1-1.3 everywhere except in Asuncion and its surrounding regions where it varies from 1.3-1.4 (Figure 46a). From 1965-2004, the PEVI increases in the major parts but it is increasing more rapidly in Asuncion and its surrounding areas as compared to the other parts (Figure 46b). The number of extremes occurring consecutively for 2 days is less than 10% of the total extremes whereas the major parts of the country do not receive consecutive 3-days extremes (Figure 47a and 47b). Decreasing trends in both consecutive 2- and 3-days extremes are observed from 1965-2004 (Figure 47c and 47d). In Asuncion, less than 5% and 2% of the total extremes occur for 2 and 3 days consecutively, respectively, and their trends show decreasing behavior from 1965-2004. Paraguay receives heavy rainfall in summer, i.e., October to March, and experiences the most number of extremes from December to February with December receiving the highest number of extremes (Figure 48).

5.3.6 Suriname and French Guiana

In Suriname and French Guiana, thresholds are low but their trends from 1965-2004 show sharply increasing behavior everywhere including Paramaribo and Cayenne (Figure 43). No variations in 50-year and 200-year RLs are observed but trends in 50-year and 200-year RLs decrease in Suriname and increase in some parts of French Guiana including Cayenne (Figures 44 and 45). The PEVI lies between 1.2 and 1.4 in Suriname whereas it ranges from 1.2-1.3 in French Guiana (Figure 46a). From 1965-2004, the PEVI shows decreasing trends in both Suriname and French Guiana except Cayenne where the PEVI increases (Figure 46b). The number of consecutive 2-days extremes varies from 1-20% in Suriname with Paramaribo receiving 15-20%

whereas it ranges from 10-15% in French Guiana (Figure 47a). Suriname receives less than 4% of the total extremes for 3-days consecutively with Paramaribo receiving 2-4% whereas in French Guiana, the number of consecutive 3-days extremes is less than 2% (Figure 47b). From 1965-2004, both consecutive 2- and 3-days extremes show decreasing trends in Suriname and increasing trends in the major parts of French Guiana including Cayenne (Figures 47c and 47d). Suriname receives heavy rainfall from April to August and experiences most of the extremes from April to June with May being the most critical with respect to the number of extremes (Figure 48). In French Guiana, the rainy season goes from April to July and the most intense months in terms of the number of extremes are from March to May with May receiving the highest number of extremes (Figure 48).

5.3.7 Extremes with topography and vegetation

In mid- and high-altitudes of the Brazilian highlands and east Venezuela, the PEVI varies from 1.1-1.4 and shows increasing trends from 1965-2004 in some areas (Figure 46a). The high-altitudes of the Brazilian Highlands receive 20-35% and 6-16% of the total extremes consecutively for 2 and 3 days, respectively, and their trends show increasing behavior from 1965-2004 in some parts (Figure 47). But in the high-altitudes of east Venezuela, consecutive 3-days extremes are not observed and less than 5% of the total extremes occur consecutively for 2 days and its trends show decreasing behavior from 1965-2004. The mid-altitudes and lowlands of east Brazil indicate very less variations in PEVI which lies between 1.1 and 1.2 but most of the eastern Brazil, which includes Rio De Janeiro and São Paulo, experience increasing trends in PEVI from 1965-2004 (Figure 46). In east Brazil, the number of consecutive 2- and 3-days extremes range from 10-30% and 1-14%, respectively, and their trends show increasing behavior from 1965-2004 only in some parts including São Paulo (Figure 47). Catingas with lowlands has the highest PEVI in South America whereas the lowlands of the Amazon basin and the Mato Grasso Plateau have higher PEVI values (Figure 46a). In the lowlands of Venezuela, north Argentina, Uruguay, Paraguay, Suriname, and French Guiana, the PEVI lies between 1.1 and 1.3 (Figure 46a). From 1965-2004, the major parts of the lowlands of the Amazon basin, north Venezuela, north Argentina, Uruguay, and Paraguay experience increasing trends in PEVI whereas decreasing trends in PEVI are observed in the lowlands of the Mato Grasso Plateau, Suriname, and French Guiana except Cayenne (Figure 46b). All the lowlands regions of Brazil, north Argentina, Venezuela, Paraguay, Uruguay, Suriname and French Guiana experience less than 10% of the total extremes for 2 days consecutively whereas the number of consecutive 3-days extremes is zero in most of these areas (Figures 47a and 47b). Only some of the lowlands areas of the Amazon basin particularly western parts of the basin, the Mato

Grasso Plateau, south Venezuela, and north Argentina experience increasing number of consecutive 2-days extremes from 1965-2004 (Figure 47c).

In the evergreen forests of the Amazon basin, south Venezuela, Suriname, and French Guiana, the PEVI ranges from 1.1-1.3 (Figure 46a). Catingas with evergreen forest has the highest PEVI in South America. Increasing trends in PEVI are observed only in some parts of the Amazon basin and south Venezuela (Figure 46b). In the Amazon basin, Catingas, and south Venezuela, the number of consecutive 2-days extremes is less than 10% of the total extremes and its trends show increasing behavior from 1965-2004 (Figure 47a and 47c). In Suriname, the number of consecutive 2-days extremes varies from 1-20% and shows decreasing trends from 1965-2004 whereas it ranges from 10-15% and indicates increasing trends from 1965-2004 in French Guiana. The number of consecutive 3-days extremes is zero in the evergreen forests of Catingas, south Venezuela, and some parts of the Amazon basin and shows decreasing trends in Suriname and increasing trends in the major parts of French Guiana from 1965-2004 (Figure 47b and 47d). In the savannas of north Venezuela, the PEVI ranges from 1.2-1.6 and shows increasing trends from 1965-2004 while the number of consecutive 2-days extremes is less than 10% and shows decreasing trends from 1965-2004 (Figures 46 and 47). In the cropland/natural vegetation of the Brazilian highlands, the Mato Grasso Plateau, east Brazil, north Argentina, Uruguay, and Paraguay, the PEVI is low and lies between 1.1 and 1.3 and shows increasing trends from 1965-2004 in some of their areas (Figure 46). The Brazilian highlands and NE Brazil experience 20-35% and 6-16% of the total extremes consecutively for 2 and 3 days, respectively, and their trends show increasing behavior from 1965-2004 in some areas (Figure 47). Less than 10% of the total extremes occur consecutively for 2 days in the Mato Grasso Plateau, SE Brazil, north Argentina, Uruguay, and Paraguay but their trends increase from 1965-2004 in some parts of the Mato Grasso Plateau and north Argentina. The number of consecutive 3-days extremes is less than 2% in SE Brazil, some parts of the Mato Grasso Plateau, north Argentina, Uruguay, and Paraguay. From 1965-2004, trends in consecutive 3-days extremes decrease in SE Brazil, the Mato Grasso Plateau, Uruguay, Paraguay but increase in some parts of north Argentina.

A caution should be exercised while interpreting the results at all those grid points where $\overline{D}_{\mathrm{SP}}$ is greater than one since at these grid points, we reject with 95% confidence that the inter-arrival times of threshold excesses follow a *homogeneous Poisson process*. The variability of extremes needs to be interpreted with care in view of issues like spatio-temporal variability in the quality of the observations as well as the possible influence of geographical features, atmospheric conditions, climate teleconnections and other phenomena that have not been considered in this dissertation. The insights on the spatial and temporal variability of extremes

will probably be more relevant in a comparative sense and at aggregate space-time scales rather than for the extremal analysis of individual points or for understanding localized phenomena related to extremes.

5.4 Summary and conclusions

This dissertation analyzed the spatial and temporal variability of precipitation extremes in South America based on daily precipitation data available in 2.5^0 gridded fields from 1940-2004. At each grid point, 65 years of data from 1940-2004 were used to understand spatial variability whereas the temporal variability was investigated for 40 years (1965-2004) and was given as the slope of linear trend obtained by fitting a regression line to 16 values computed from 25-year moving window from 1965-2004, i.e., 1965-1989, 1966-1990,. . ., 1980-2004. We analyzed weekly precipitation maxima residuals and utilized the Poisson-GP model to investigate the spatial and temporal variability of threshold, the scale (σ) and shape (ξ) parameters, and 50-year and 200-year RLs. The temporal variability of precipitation were evaluated from the temporal variability of thresholds. The threshold was chosen as the 95%-quantile of time series. We also investigated the spatial and temporal variability of the PEVI, which measures the variability of extremes and is defined as the ratio of 200-year and 50-year RLs. Based on daily precipitation data, we investigated the spatial and temporal variability of the percentage of the number of consecutive 2- and 3-days extremes out of the total number of extremes. The spatial variability of the percentage of the number of extremes in a particular month out of the total number of extremes was also investigated based on daily precipitation data. The threshold for the analysis of daily precipitation data was chosen as the 99%-quantile of time series.

Precipitation is high indicated by high thresholds in SE Brazil, Uruguay, Paraguay, and Buenos Aires. The PEVI is high in the eastern parts of the Amazon basin, Catingas, Mato Grasso Plateau, NW Venezuela including Caracas, and Asuncion. From 1965-2004, both precipitation and the PEVI show increasing trends in the eastern coastal regions of Brazil including Rio De Janeiro, the Brazilian highlands particularly southern parts, north Venezuela including Caracas, some parts of north Argentina, Uruguay, Paraguay including Asuncion, and Cayenne. The Amazon basin except eastern parts and São Paulo experience increasing trends in the PEVI and decreasing trends in precipitation. In the eastern parts of the Amazon basin, Catingas, the Mato Grasso Plateau, Brasilia, Buenos Aires, and Montevideo, simultaneous decreasing trends are observed in precipitation and the PEVI. The PEVI shows decreasing trends in Suriname including Paramaribo and French Guiana excluding Cayenne although increasing precipitation trends are observed in these areas. The number of consecutive 2- and 3-days extremes are high in the Brazilian Highlands, NE Brazil, and Brasilia. Trends in precipitation and the number of both consecutive 2- and 3-days extremes increase in few parts of

the Brazilian Highlands and some parts of French Guiana including Cayenne. Catingas and some parts of east Brazil including São Paulo experience increasing trends in the number of consecutive 2- and 3-days extremes and decreasing trends in precipitation. The precipitation shows increasing trends whereas the number of consecutive 2- and 3-days extremes show decreasing trends in NE coastal regions of Brazil, Rio De Janeiro, Uruguay, Paraguay including Asuncion, and Suriname including Paramaribo. In Brasilia, both precipitation and the number of consecutive 2- and 3-days extremes show decreasing trends simultaneously. The number of consecutive 2-days extremes also show increasing trends in the Amazon basin particularly western parts, some parts of the Mato Grasso Plateau, and Buenos Aires although precipitation shows decreasing trends in these areas. Simultaneous decreasing trends are observed in precipitation and the number of consecutive 3-days extremes in the Mato Grasso Plateau and Buenos Aires from 1965-2004.

The areas of interest based on an increasing PEVI from 1965-2004, are the Amazon basin, the Brazilian Highlands, Venezuela, Uruguay, Paraguay, and some of the highly populated cities in South America, specifically Rio De Janeiro, São Paulo, Caracas, Asuncion, and Cayenne. Some parts of east Brazil, few parts of the Brazilian highlands, São Paulo, and Cayenne also experience increasing number of consecutive 2- and 3-days extremes. Water resources engineers and planners, disaster management agencies, and policy makers need to pay special attention to the regions with increasing trends in the PEVI and consecutive 2- and 3-days daily extremes, especially when these regions overlap with densely populated areas, while planning for infrastructure development and disaster management. Civil engineers can utilize the results of this dissertation for the design of hydraulic structures, specifically when considering the optimal safety factors in their design. Hydrologists and climatologists need to delve deeper into the potential causes of the observed spatio-temporal trends in extremes for delineating the variability of extremes due to natural and anthropogenic effects.

Precipitation extremes may result in significant loss of human life and property. However, the damages caused by precipitation can be influenced by a variety of factors other than just precipitation maxima or the statistical properties thereof. These factors include surface and sub-surface hydrology since the damages caused by precipitation extremes are primarily caused by floods and flash floods, which, in turn, are strongly influenced by the physics of runoff and infiltration. The other factor is population: certainly the (catastrophic) impact of disasters depends on (high) population densities and the location of critical infrastructures or national/human assets which may be potentially damaged by precipitation extremes. Finally, the actual damages would also be a function of resilience of communities and critical infrastructures to precipitation extremes and related disasters.

Future research needs to explore the use of extreme value theory in conjunction with more advanced physically-based or statistical models of precipitation, as well as the utilization of emerging techniques for the estimation of the extreme value parameters directly as a function of time, seasonality and other covariates. In addition, development of heuristic approaches for the estimation of optimal thresholds in the context of precipitation data may need to be explored, since similar approaches developed for other types of data may or may not be directly applicable to precipitation data, especially when the data sets are large. Hydrologists and climatologists can perform further research based on this dissertation to understand the natural or anthropogenic causes driving precipitation extremes and their spatial or temporal trends. Future research may combine the PEVI used in this dissertation with other factors like population and critical infrastructures to estimate the potential risks from extremes and subsequently with development or financial indices to estimate the corresponding impacts.

Acknowledgements

This research was funded by the SEED of the Laboratory Directed Research and Development (LDRD) Program of the Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract DE-AC05- 00OR22725 (Project Title: "Multivariate Dependence in Climate Extremes"; PI: Auroop R. Ganguly). We would like to gratefully acknowledge Dave Allured and Brant Liebmann from the NOAA-CIRES Climate Diagnostics Center, Boulder, Colorado, for providing us the dataset used in this dissertation. The authors are thankful to two anonymous reviewers for their helpful suggestions which significantly improved the quality of the paper. Shiraj Khan would like to acknowledge the help and support provided by Prof. Sunil Saigal of the Civil and Environmental Engineering at the University of South Florida. Auroop R. Ganguly would like to thank Dr. Rick Katz of the National Center for Atmospheric Research and Prof. Tailen Hsing of the Ohio State University for their help and support with the SEED project. The bulk of this work was completed while Shiraj Khan and Gabriel Kuhn were post-graduate employees at ORNL. This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government and the publisher, by accepting the article for publication, acknowledge that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for its purposes.

Chapter 6

Detection and Predictive Modeling of Chaos in Finite Hydrological Time Series

The ability to detect the chaotic signal from a finite time series observation of hydrologic systems is addressed in this paper. The presence of random and seasonal components in hydrological time series, like rainfall or runoff, makes the detection process challenging. Tests with simulated data demonstrate the presence of thresholds, in terms of noise to chaotic-signal and seasonality to chaotic-signal ratios, beyond which the set of currently available tools is not able to detect the chaotic component. The investigations also indicate that the decomposition of a simulated time series into the corresponding random, seasonal and chaotic components is possible from finite data. Real streamflow data from the Arkansas and Colorado rivers are used to validate these results. Neither of the raw time series exhibits chaos. While a chaotic component can be extracted from the Arkansas data, such a component is either not present or can not be extracted from the Colorado data. This indicates that real hydrologic data may or may not have a detectable chaotic component. The strengths and limitations of the existing set of tools for the detection and modeling of chaos are also studied.

6.1 Introduction

The presence of nonlinear dynamics and chaos has strong implications for predictive modeling and the analysis of dominant processes in any discipline. The existence of chaotic behavior has been demonstrated in diverse areas ranging from turbulence [119], weather or climate [67, 120–122] and geophysics [54, 123–126], to biology or medicine [127], finance [128–130], and electrical circuits [131]. The presence of chaos in hydrology has been suggested by previous researchers [48–64]. The ability to detect and model chaotic behavior from finite hydrologic time series has recently been debated [65, 66].

Characterization of chaos from real-world observations is known to be a difficult problem in nonlinear dynamics [67–69]. The complexity was highlighted in the context of climate models by [70], who demonstrated that sensitivity to initial conditions may become less apparent when the randomness in internal atmospheric variables begins to dominate. Fundamental questions still remain unanswered in these areas, for example the ability to detect chaos from a finite time series with random and seasonal components, the ability to

decompose a time series into these components, and the corresponding implications for predictive modeling. However, addressing these questions is critical for hydrology. This can be gauged from the wealth of hydrologic literature in areas like complexity analysis [71–73], predictability [64] and nonlinear predictive modeling [74–77].

This paper investigates the ability of nonlinear dynamical tools to detect, characterize, and predict chaos from finite hydrologic observations, using both simulated and real time series. Realistic simulated data is generated by contaminating chaotic signals with random and seasonal components, while real streamflow data are used from the Arkansas and Colorado rivers. The correlation dimension method is used for detecting the possible presence of chaos. Nonlinear predictive models, namely the phase-space reconstruction (PSR) and artificial neural networks (ANN) are employed for time series decomposition and prediction. This dissertation develops several new insights and interesting results. The presence of thresholds for the detect-ability of chaos is demonstrated, specifically when a chaotic signal is mixed with random or seasonal signals, or a combination thereof. These thresholds can be expressed in terms of the relative dominance of the chaotic component compared to the random or seasonal components. The ability to decompose a time series into the contributions from the individual components (random, seasonal and chaotic) is shown. The corresponding implications for predictive modeling and characterizing the nonlinear dynamics are highlighted. Real streamflow data analysis provides additional insights. First, not all hydrologic time series contain chaotic components which can de detected and modeled. Second, for certain finite hydrologic time series, the presence of chaos can indeed be detected, isolated from random and seasonal components, and utilized for predictive modeling.

The rest of the paper is organized as follows. Section 2 presents the tools and methods employed in this dissertation. The simulated and real hydrological data are discussed in Section 3. Sections 4 and 5 present and discuss the results obtained with simulated and real hydrological time series, respectively. The summary and conclusions of this paper are presented in Section 6.

6.2 Tools and methods

6.2.1 State of the art and literature review: tools and concepts

The theoretical concepts underlying the methodologies for the detection and modeling of nonlinear dynamical and chaotic components, as well as their implementation, are available in the literature [64, 132–134]. The present dissertation exploits the correlation dimension method, as well as nonlinear predictive models like

PSR and ANN. These tools have been used widely for the identification and modeling of nonlinear dynamics and low-dimensional chaos in short hydrological time series [50, 59, 64]. Contributions include refinement of these tools in terms of either the underlying methodologies or the applications. A brief discussion of the tools and the present modifications is described in this section.

6.2.1.1 Correlation dimension An algorithm suggested by [135] is the most commonly used method in hydrology to characterize chaotic attractors. In a time series of a single dynamic variable, z_i, where $i = 1, 2, .., N$, a vector in an m-dimensional phase space may be given as $\mathbf{Z}_j = (z_j, z_{j-\tau}, .., z_{j-(m-1)\tau})$, where $j = \frac{m\tau}{\Delta t}, .., N$; Δt is the time interval; τ is the delay time; and m is the embedding dimension of the state space. The correlation sum $C(r)$ is expressed as

$$C(r) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{i-1} \Theta(r - \|\mathbf{Z}_i - \mathbf{Z}_j\|) \tag{36}$$

where Θ is the Heaviside step function; N is the number of points in the time series; and r is the radius of a sphere with its center at either of the current points, z_i or z_j. The relation between correlation sum $C(r)$ given by Eq. (1) and correlation exponent ν is expressed as

$$C(r) = \lim_{r \to \infty} c r^{\nu} \tag{37}$$

where c is a constant and $\nu = \lim_{r \to \infty} \frac{Ln\ C(r)}{Ln\ r}$. Since a real data set consists of a finite number of points, there always exists a minimum distance, d_{min}, between the trajectory points. When $r < d_{min}$, the correlation sum $C(r)$ is zero and no longer scales with r. Therefore, Eq. (2) with limit $r \to \infty$ cannot be used directly to determine the correlation exponent. In an alternative technique, a plot of $Ln\ C(r)$ vs. $Ln\ r$ graph is obtained. Several methods are available to determine the correlation exponent from this plot. The slope of the *scaling region*, where the curve can be approximated as a straight line, gives the correlation exponent ν. In the present dissertation, the *least squares* method is used to fit a line using a moving window of size d to plot the *slope* vs. $Ln\ r$ curve. The adjacent slope values lying close to a horizontal portion of the curve are averaged to determine the correlation exponent. For the system to be chaotic, the correlation exponent should increase up to a certain point and then saturate with an increase in the embedding dimension. To calculate the saturation value of the correlation exponent, the acceptable error is defined as $\varepsilon = p*(max_{ci} - min_{ci})$, where p is an acceptable percentage, and max_{ci} and min_{ci} are the respective maximum and minimum correlation exponents in a set of M values. In this dissertation, the values of p and M are taken as 1% and 20, respectively. The difference between the adjacent values of the correlation exponent is given as $\delta = C_j - C_i$, where

$j = i + 1$. If $\delta \leq \varepsilon$, then $\Delta = \frac{|C_M - C_i|}{M-i}$. If $\Delta \leq \varepsilon$, then C_i is the saturation value of the correlation exponent. The nearest integer above the saturation value of the correlation exponent gives the correlation dimension of the attractor. It has been proposed [59, 136] that the correlation dimension may be related to the minimum number of variables required to extract a multidimensional description of the dynamical system. If $\delta > \varepsilon$ and saturation never occurs, the presence of chaos cannot be confirmed.

[137] observed that the finite correlation exponent achieved using the correlation dimension method is not a good indicator of the presence of chaos since linear stochastic processes may also yield a finite correlation exponent. The determination of correlation exponent is greatly influenced by several factors including limited data size, the presence of noise, delay time, and the presence of a large number of zeros in the data set. Hydrological time series is finite, contaminated with noise, and may contain a large number of zeros. A finite and small data set produces a smaller scaling region. This may not be sufficient to calculate the slope of the $Ln\ C(r)$ vs. $Ln\ r$ curve and may result in an underestimation of the correlation exponent. A large scaling region may be better delineated when a large data set is used which, in turn, results in a better estimation of the slope. [138] suggested that the minimum number of data points required for the correct estimation of the correlation exponent is $N_{min} = 10^{2+0.4m}$, where m is the embedding dimension. The presence of noise may affect the scaling behavior and may tend to make the slope of the $Ln\ C(r)$ vs. $Ln\ r$ plot larger for small values of r resulting in an overestimation of the correlation exponent. If the delay time, τ, is too small, the phase space may contain very little information and may result in an underestimation of the correlation exponent. If τ is too large, the phase space may miss out nearby diverging trajectories resulting in an overestimation of the correlation exponent. The presence of a large number of zeros in the time series produces a phase space with limited information about the underlying dynamics and results in an underestimation of the correlation exponent.

6.2.1.2 Artificial neural networks (ANNs) ANNs have been used widely for modeling nonlinear hydrologic processes such as rainfall-runoff, streamflow, ground-water management, water quality simulation, and precipitation [139]. A prediction is expressed by a function, $\hat{\mathbf{f}}_{\mathbf{NN}}$, that maps the input $(z_t, z_{t-\tau}, .., z_{t-(m-1)\tau})$ to the expected output $\hat{z}_{t+T}$, i.e. $\hat{z}_{t+T} = \hat{\mathbf{f}}_{\mathbf{NN}}(z_t, z_{t-\tau}, .., z_{t-(m-1)\tau})$, where τ and m are the delay time and number of variables required to model the system, respectively. The present dissertation employs the multi-layer perceptron (MLP) as a nonlinear function approximator in the context of time series observations [140, 141]. For a m-dimensional phase space, an appropriate ANN architecture could comprise m inputs and one output. The number of hidden layers and corresponding hidden nodes would then depend on the complexity of the functional form to be modeled. In practical applications, more than two hidden

layers are seldom used, while a single hidden layer is usually deemed adequate. The choice of the optimal number of hidden nodes is still an open research topic, although methods based on information criteria and other approaches have been proposed [142]. For the purposes of this dissertation, a single hidden layer with $integer(m/2)$ hidden nodes is employed. The ANN is trained using the commonly used backpropagation algorithm. This algorithm repeatedly computes an error between the output of the network and the desired output and feeds this error back to the network. The error input is used to adjust the weights until the error is minimized. In this dissertation, two-third of the data set is used for training and the remaining one-third is used for predictions. The accuracy of the prediction is calculated in terms of the goodness of fit statistics, *correlation coefficient (CC)*, and error statistics, *mean squared error (MSE)*, between the original data and the predicted data.

As the number of input nodes in the ANN is increased, the prediction skills increase up to a certain point and then become constant. A large number of input nodes reflects the use of additional lagged variables to model the time series. When the number of inputs is too low, the information content in the lagged values is not adequately captured. This results in a higher bias and hence lower prediction skills. When the number is too large, the functional form to be modeled grows complex, leading to a larger error variance and corresponding decay in skills. This bias-variance tradeoff (and related issues) usually results in an optimal number of nodes where the skills attain a maxima. In certain cases, the number of input values representing the number of lagged variables of a time series at which the skills attain a maxima have been equated with the number of variables influencing the dynamical system (e.g., [59]).

6.2.1.3 Phase-space reconstruction (PSR) prediction The PSR [143] has been used widely for predicting chaotic time series in hydrology. The time series with a single dynamical variable is embedded in a m-dimensional state space. The dynamical system can be interpreted in the form of a nonlinear function $\mathbf{F}_T$, i.e. $\mathbf{Z}_{i+T} = \mathbf{F}_T\mathbf{Z}_i$, where $\mathbf{Z}_i$ and $\mathbf{Z}_{i+T}$ are the current state space and the future state space, respectively. In order to predict $\hat{z}(i+T)$ using the current state $\mathbf{Z}_i$, a function $\hat{\mathbf{F}}_T$ that maps a m-dimensional state space to a single scalar variable of the system is employed as

$$\hat{z}_{i+T} = \hat{\mathbf{F}}_T\mathbf{Z}_i \tag{38}$$

In the present dissertation, the *local approximation* approach suggested by [143] is used to determine $\hat{\mathbf{F}}_T$. This approach reduces the complexity of this function by dividing its domain into many neighborhoods and estimating an approximation map for each neighborhood. As suggested by [126], $\hat{\mathbf{F}}_T$ can be modeled as a

linear function for best results. Two-third of the data set is used for calibrations and the remaining one-third for predictions. The function of Eq. (3) is modeled by using a linear regression model as

$$\begin{bmatrix} z_{11+T} \\ z_{22+T} \\ \cdots \\ z_{kk+T} \end{bmatrix} = \begin{bmatrix} z_{11} & z_{11-\tau} & \cdots & z_{11-(m-1)\tau} \\ z_{22} & z_{22-\tau} & \cdots & z_{22-(m-1)\tau} \\ \cdots & \cdots & \cdots & \cdots \\ z_{kk} & z_{kk-\tau} & \cdots & z_{kk-(m-1)\tau} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_m \end{bmatrix} \tag{39}$$

where, $(z_{11}, .., z_{11-(m-1)\tau}), .., (z_{kk}, .., z_{kk-(m-1)\tau})$ are the k states nearest to the current state, $\mathbf{Z}_i$, and are computed based on the Euclidean distance measure $\|\mathbf{Z}_i - \mathbf{Z}_j\|$, where $\mathbf{Z}_j = (z_{jj}, .., z_{jj-(m-1)\tau})$; $b_1, .., b_m$ are the function coefficients; and $z_{11+T}, .., z_{kk+T}$ are the k predicted data points. Using function coefficients, b_i, from Eq. (4), the predicted value for the current state is given as $\hat{z}_{i+T} = \sum_{r=1}^{m} z_{i-(r-1)\tau} b_r$, where τ is the delay time. In the present dissertation, the values employed for k and τ are 25 and 1, respectively. The prediction accuracy is represented in terms of CC and MSE between the original data and the predicted data. It has been suggested [59] that the embedding dimension at which the *CC* vs. *m* curve saturates may indicate the minimum number of variables required to capture the dynamics of the system.

6.3 Data description

The effectiveness of the methods described in section 2.1 in terms of their ability to detect the presence of low-dimensional chaos, as well as for short-term predictions, is investigated. This is accomplished through an analysis of both simulated and real hydrologic data.

6.3.1 Simulated data

Hydrological time series are always finite and may be contaminated with noise and seasonality. To obtain realistic insights, simulated data is generated by contaminating chaotic time series with noise, i.e. white and autoregressive, and seasonality. The chaotic time series is represented here by the Lorenz system of equations as

$$\frac{dx}{dt} = \beta(y - x), \frac{dy}{dt} = -xz + rx - y, \frac{dz}{dt} = xy - bz \tag{40}$$

where $\beta = 10$; $r = 28$; and $b = 8/3$. The Lorenz system is highly sensitive to the initial conditions and is, therefore, chaotic. Seasonality implies periodicity which, in turn, could broadly include

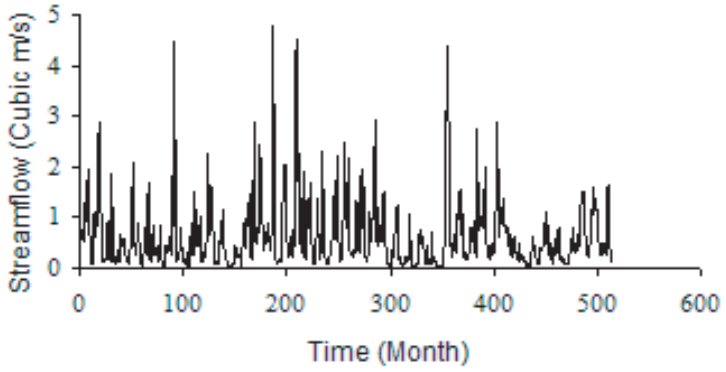

Figure 50. Monthly streamflow time series observed at the Arkansas river.

intra-annual and inter-annual cycles. For this dissertation, seasonality is represented by a periodic function, $x(t) = A\ cos(2\pi ft)$, where A and f are the amplitude and frequency, respectively. Noise consists of random variations in the data. White noise is generated by using normal distribution with $\mu(mean) = 0$ and a user specified value of $\sigma(standard\ deviation)$.

6.3.2 Hydrologic time series

Real hydrologic series including monthly streamflow time series of the Arkansas river at Little Rock and daily streamflow time series of the Colorado river below Parker dam are investigated. The raw data is obtained from the *U.S. Geological Survey* site.

The presence (or absence) of low-dimensional chaos in a short hydrological (monthly streamflow) time series at the Arkansas river at Little Rock in Arkansas is first studied. The Arkansas River is the fourth longest river in the United States and is a tributary of the Mississippi which flows east and southeast through Colorado, Kansas, Oklahoma and the state of Arkansas. It is located at Latitude $34^o45'00''$ and Longitude $92^o16'25''$. The associated drainage area is 158,090 square miles while the contributing drainage area is 135,849 square miles. The temperature at Little Rock ranges from a mean low of 40^oF in January to a mean high of 81^oF in July. The mean annual precipitation at Little Rock is 50.26 inches. The streamflow data of 43 years (October 1927 - September 1970) is analyzed. The monthly streamflow time series at the Arkansas river is shown in Fig. 50.

The daily streamflow data observed at the Colorado river below Parker dam, Arizona-California is also investigated for the existence of chaotic behavior. The Colorado River flows through Colorado, New Mexico, Utah, California, Arizona and Nevada. It drains a part of the arid regions on the western slopes of the Rocky

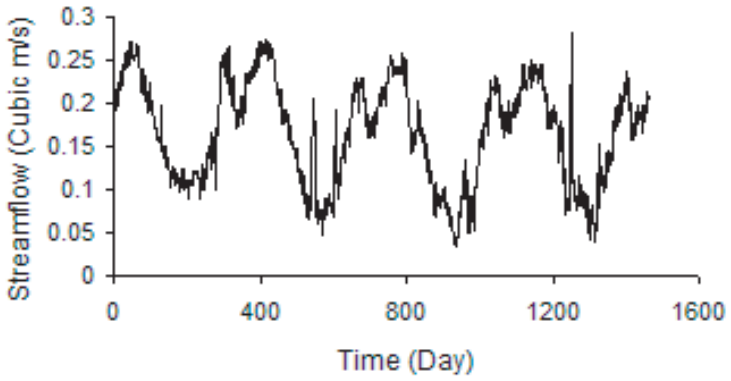

Figure 51. Daily streamflow series observed at the Colorado river.

Mountains. The height, crest length and base thickness of Parker dam are 320 feet, 856 feet, and 100 feet, respectively. It is located at Latitude $34^o17'44''$, Longitude $114^o08'22''$. The associated drainage area is 182,700 square miles while the contributing drainage area is 178,700 square miles. The average temperature varies from 43.1^oF in winter to 107.5^oF in summer. The annual average precipitation is 6.2 inches. The daily streamflow data observed for 4 years (June 1, 1959 - May 31, 1963) is analyzed. The variation of the daily streamflow series at the Colorado river is shown in Fig. 51.

Statistics of the streamflow data observed at the Arkansas and Colorado rivers are given in Table 12.

Table 12. Streamflow data statistics (values in m^3/s)

Parameter	Arkansas river	Colorado river
Data points	516	1461
Number of zeros	0	0
Mean	0.6529	0.1672
Standard deviation	0.684	0.0594
Variance	0.4678	0.00353
Maximum value	4.7571	0.2819
Minimum value	0.0187	0.0334

6.4 Results with simulated data

6.4.1 Pure chaotic, random and seasonal time series

In this dissertation, a chaotic time series is represented by the X component of Eq. (5); a random series is a normally-distributed white noise with $\mu = 0$ and $\sigma = 0.16$; and seasonal series is represented by a periodic function, i.e. cosine function with a frequency of $10Hz$ and an amplitude 10. Each series has 1000

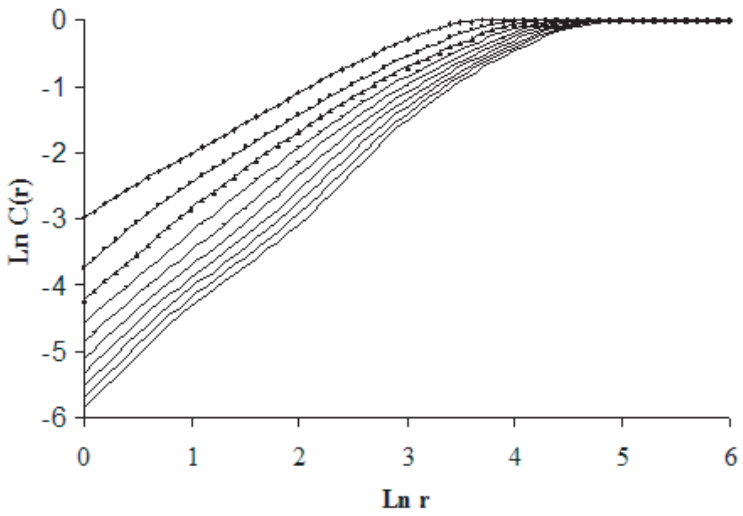

Figure 52. $LnC(r)$ vs. Lnr plot for Lorenz (X component) time series. The curves are shown from top to bottom in ascending order of embedding dimension, m = 2, 4,.., 20.

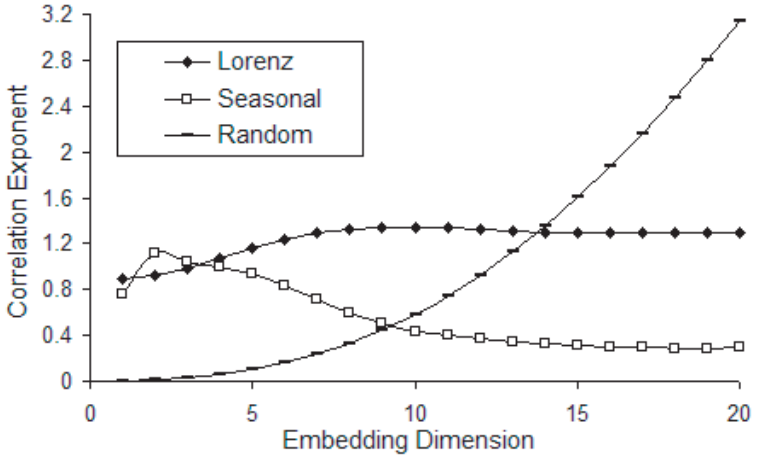

Figure 53. Relation between correlation exponent and embedding dimension for Lorenz (X component), seasonal, and white noise series.

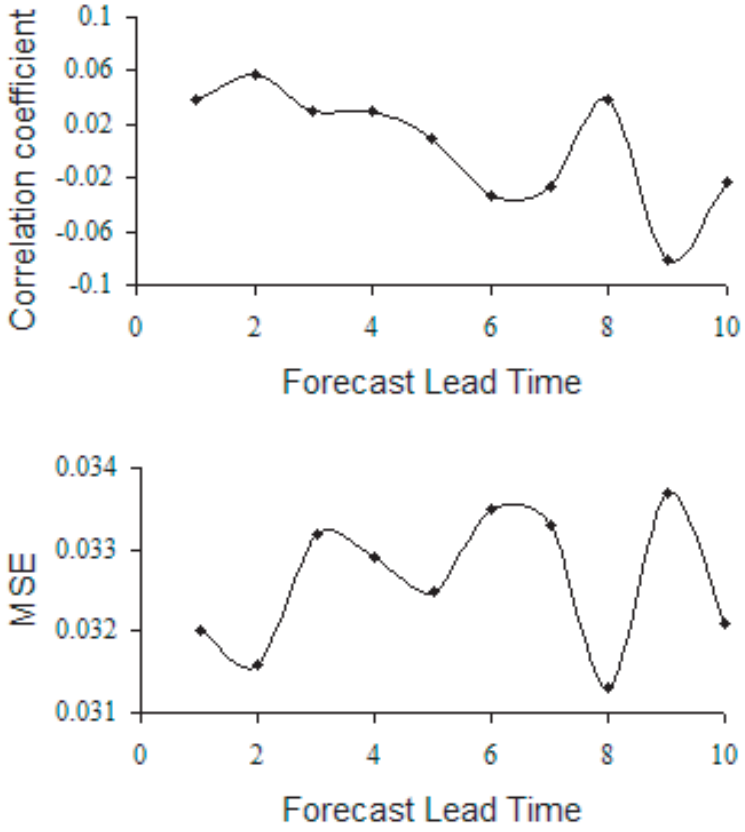

Figure 54. The variation of CC and MSE with forecast lead time for white noise with $\sigma = 0.16$.

data points, which is greater than the number of points required ($10^{2+0.4*m} = 650$, where $m = 2$ for the Lorenz time series) for the correct estimation of the correlation exponent of the Lorenz time series [138]. The plot of $Ln\ C(r)$ vs. $Ln\ r$ described in section 2.1.1 for Lorenz time series is shown in Fig. 52, while the corresponding $correlation\ exponent$ vs. m plot is shown in Fig. 53. For chaotic series, the correlation exponent saturates at an embedding dimension of 10. The saturation value of the correlation exponent is 1.34, while the correlation dimension of the Lorenz attractor is 2. For random series, no saturation is observed as seen in Fig. 53. The correlation exponent shown in Fig. 53 decreases rather markedly for the seasonal series considered here, while the $Ln\ C(r)$ vs. $Ln\ r$ plot shows characteristic steps [134]. For purposes of generality, uniform and objective methods are utilized to identify the scaling regions and to calculate the slope through the moving window technique described above. A window size of 9 is chosen and the average of ten slope values from $Ln\ r = 0.8$ to $Ln\ r = 1.7$ in increments of 0.1 are taken. These parameters are selected through trial and error. Once selected, they are kept constant for all the case studies with simulated data. While this ensures uniformity, there is a possibility that judgmental approaches on each series might have yielded differing values for the scaling region and smoothing windows. However, one purpose of the paper is to motivate the development of tools that can be utilized in an automated fashion and that do not rely on user judgment or subjective considerations. The results with predictive modeling approaches like PSR and ANN indicate significant prediction skills, i.e. $CC = 1$ and $MSE = 0$, for forecast lead time from one to ten at $m = 10$ for the chaotic and the seasonal series. For random series, the predictive skills fluctuate around a lower value and show only a slightly decreasing trend as shown in Fig. 54.

6.4.2 Mixed time series

As described earlier and indicated by previous researchers [60, 136, 144], hydrologic and other real systems tend to generate observables that have random and seasonal components, in addition to any nonlinear deterministic (or chaotic) signal that may be present. Simulated data are generated using a mixture of these components to understand the ability to identify, characterize and quantify chaos from amidst seasonality and noise through the commonly used tools for nonlinear dynamics and chaos.

6.4.2.1 Mixture of chaotic and seasonal series The ability to distinguish chaotic signals, when mixed with seasonal signals of varying amplitudes, is depicted by $correlation\ exponent$ vs. m plots shown in Fig. 55. These plots are obtained using the correlation dimension method. First, it is seen that as the seasonal component is increased, i.e. the amplitudes are made higher, the saturation value of the correlation exponent also increases. Second, there appears to be a threshold value, expressed as the seasonality to chaotic-signal

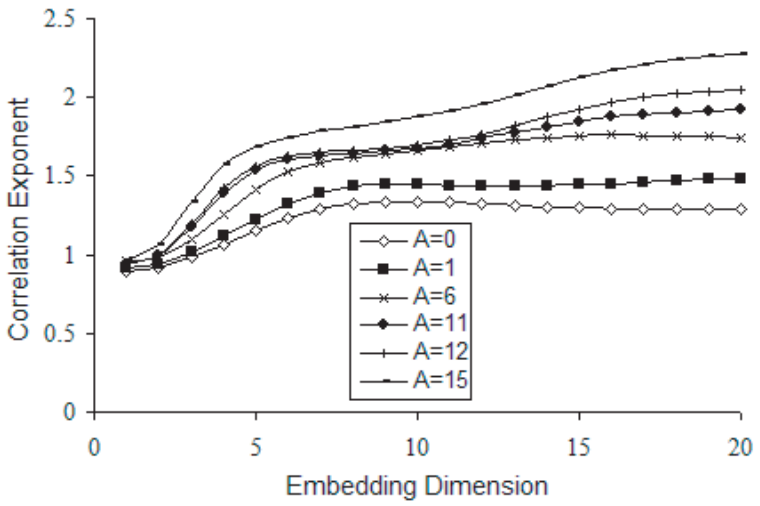

Figure 55. Relation between correlation exponent and embedding dimension for mixed time series. Series includes Lorenz X-component and seasonality with $f = 10Hz$ and different amplitudes.

ratio, for the degree of seasonality beyond which the chaotic component cannot be distinguished any longer. The degree of seasonality is expressed in terms of the amplitude of the seasonal component for fixed value of the chaotic Lorenz series. From eye estimation, for amplitudes of 1 and 6 in Fig. 55, the correlation exponent appears to saturate with an increase in the embedding dimension. Based on objective techniques established in section 2.1.1, it can be demonstrated that the saturation indeed occurs and the correlation exponent either decreases, or remains constant, or increases within tolerable limits, till an amplitude of eleven. However, for amplitudes larger than 11, the increase in correlation exponent no longer remains within the tolerable limits, thus indicating the presence of a possible threshold. While the presence of a threshold is indicated, the exact value can depend to some extent on how the criteria for saturation and the corresponding tolerances are defined. It may also depend on the number of data points used for the analysis. It is further observed that the amplitude of 11 corresponds to the highest integer value of the seasonal amplitude for which σ of the seasonal component is lower than that of the chaotic series. The σ values for the pure chaotic series, pure seasonal series with amplitude 11, and pure seasonal series with amplitude 12, are 8.31, 7.78, and 8.48, respectively. For an amplitude of 11, the seasonality to chaotic-signal ratio, i.e. $\frac{\sigma_s}{\sigma_c}$, is 0.94. Thus, the results: (a) appear to indicate the presence of a threshold value, i.e. 0.94, for the degree of seasonality beyond which chaotic components may no longer be distinguishable at least through the use of the correlation exponent based approaches, and (b) suggest that the threshold may occur at the point where the seasonality to chaotic-signal ratio roughly equals unity.

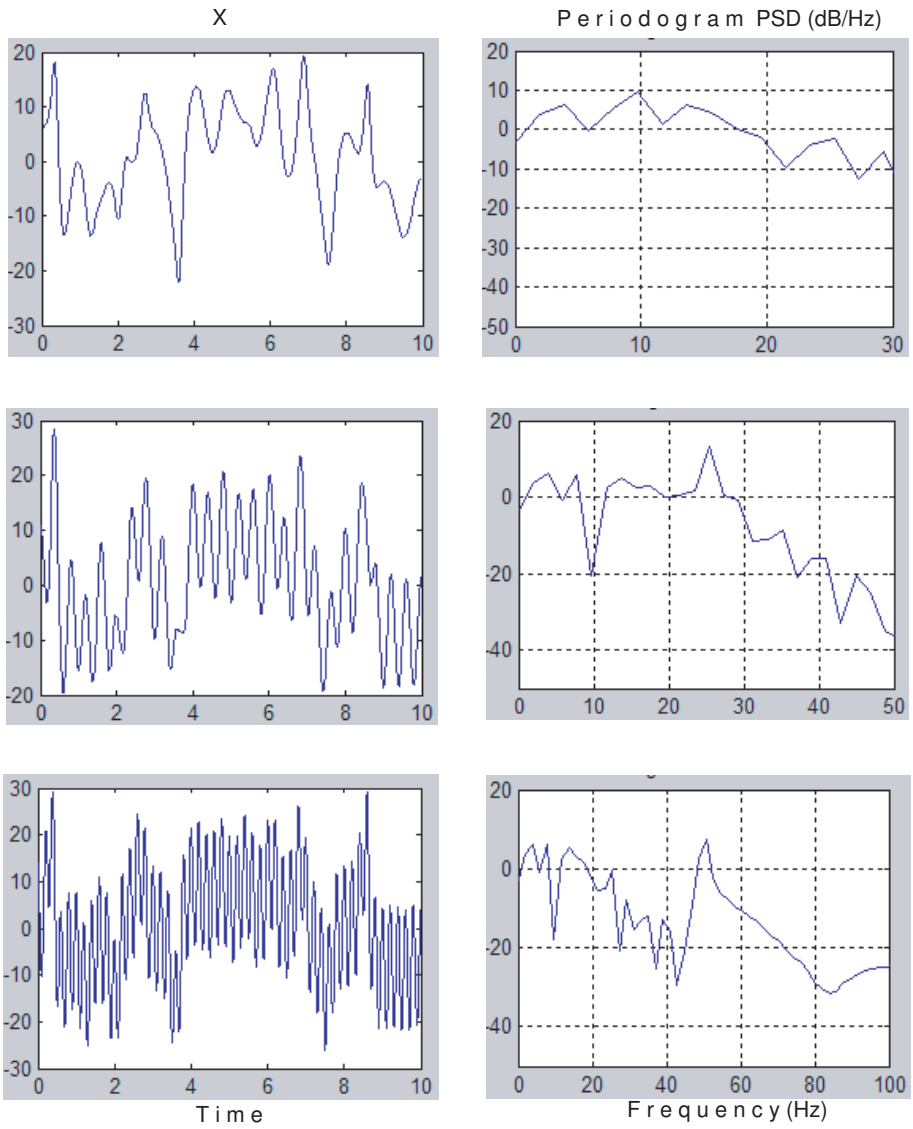

Figure 56. Mixed times series (Lorenz X-component and seasonality) and its periodograms showing the variation of power spectral density (PSD) with frequency. Top: Lorenz X-component and seasonality with $f = 10Hz$ and $A = 5$. Middle: Lorenz X-component and seasonality with $f = 25Hz$ and $A = 10$. Bottom: Lorenz X-component and seasonality with $f = 50Hz$ and $A = 13$.

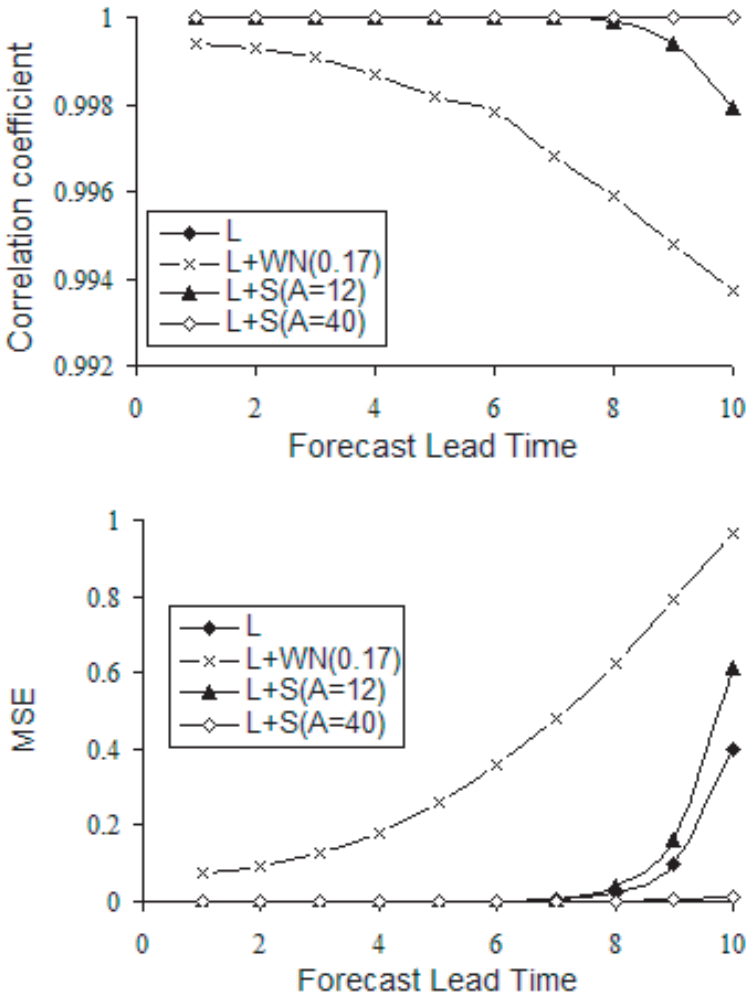

Figure 57. Variation of CC and MSE with forecast lead time for chaotic and mixed series. L, WN and S stand for Lorenz, white noise and seasonality, respectively.

The separation of the nonlinear deterministic signal from the seasonality is next considered. Frequency domain methods are applied to identify the dominant frequencies of the seasonal component. This implicitly assumes that the seasonal component, if present, in a simulated or real hydrologic series, is dominant enough to be identifiable through the periodogram estimate. Fig. 56 shows chaotic-signal mixed with seasonal signals of different frequencies and amplitudes, and the corresponding periodogram estimates. As shown in Fig. 56, the dominant frequency is observed for each simulation. The amplitudes selected here represent the minimum values of the amplitudes, corresponding to the given frequencies, above which a peak can be distinguished in the periodogram. Thus, seasonal components of higher frequencies need higher amplitudes to make them identifiable. Once the frequencies are obtained, curve fitting techniques can be utilized to obtain the amplitudes of the seasonal component. The remaining component is the non-seasonal portion, which in this case would be the chaotic component. It is noted again that when extending this observation to real data, the non-seasonal component needs to be further decomposed into random and deterministic components. The deterministic component is next analyzed for chaotic signals. The separation of random and deterministic (possibly chaotic) signals is discussed in section 4.2.2. Table 13 provides an example where the white noise component has been isolated from a series which is a mixture of chaotic signal, seasonality, and white noise. The original and recovered standard deviations of the white noise are shown along with the corresponding errors. Once the deterministic component, i.e. chaotic + seasonality, is isolated from the mixture, the seasonal component is separated by fitting a periodic curve with the dominant frequency found using frequency domain analysis. The remaining chaotic component is compared with the original Lorenz series used for the simulation. It is seen that as the white noise component in the mixture increases, the prediction skills represented by statistics *CC* and *MSE* decrease. The PSR, with the embedding dimension of 10, is used for multi-step ahead forecasting of a mixture of chaotic series and seasonality ($f = 10,\ A = 12$; and $f = 10,\ A = 40$). Fig. 57 shows the CC vs. $Forecast\ lead\ time$ and MSE vs. $Forecast\ lead\ time$ plots. Good prediction skills with $CC = 1$ and $MSE = 0$ for a mixture of chaotic and seasonal series are observed.

Table 13. Separation of white noise from a mixture of chaotic, seasonal and white noise series using the PSR with $m = 10$.

Org. σ_{wn}	Cal. σ_{wn}	% error	MSE	CC
0.075	0.0910	21.3	0.0083	0.9999
0.1	0.1254	25.4	0.0157	0.9999
0.1	0.2318	15.9	0.0537	0.9997
0.33	0.3608	9.33	0.1301	0.9992
1.0	1.1985	19.85	1.4349	0.9913
2.0	2.6403	32.01	6.9706	0.9587
3.0	4.1341	37.8	17.0917	0.9022
5.0	6.5895	31.8	43.4134	0.7806
10.0	12.624	26.24	159.3665	0.4948

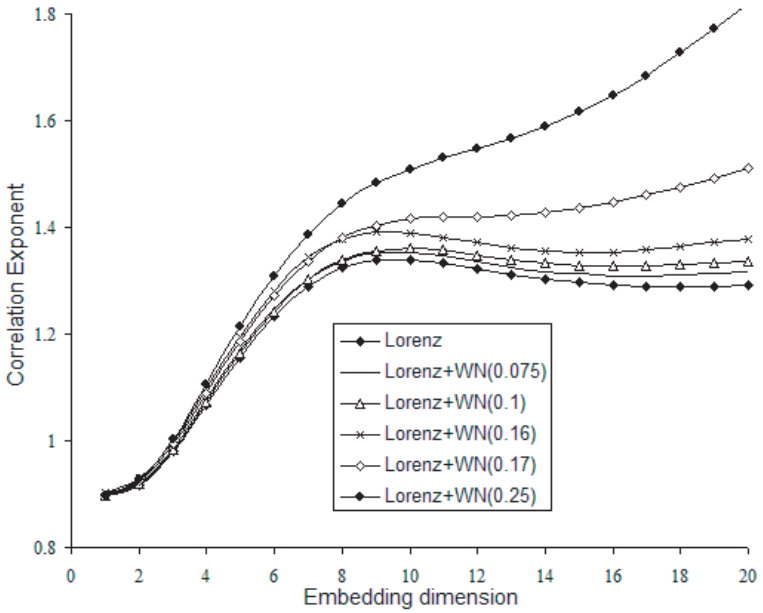

Figure 58. Correlation exponent vs. embedding dimension plot for mixed time series consisting of Lorenz (X component) and white noise.

6.4.2.2 Mixture of chaotic and random series The ability to separate the deterministic component from a mixture of chaotic and random series is studied. $Correlation\ exponent$ vs. m plot obtained using the correlation dimension method is shown in Fig. 58. Each curve represents a different value of σ for the white noise component that is mixed with the same underlying chaotic signal. Several properties are observed. First, the saturation value of the correlation exponent increases as the white noise component (σ) increases. Second, a threshold value appears to exist for the noise to chaotic-signal ratio, beyond which the latter cannot be distinguished through the use of this method. The correlation exponent does not saturate when the σ of the white noise component becomes more than 0.17, which roughly corresponds to a threshold value of $1/49$ for the noise to chaotic-signal ratio. Similar to the case of mixture of chaos and seasonality described in the previous section, the results presented here appear to suggest the existence of a threshold. The actual value of the threshold may vary depending on the predefined saturation criteria, tolerances and possibly on the number of data points. However, the results do indicate that a threshold value may exist, beyond which the chaotic component can not be determined from a mixture of chaotic-signal and white noise. The threshold value appears surprisingly low (just two percent in terms of the ratio of the standard deviations) for a chaotic-signal contaminated with pure random noise. Besides the white noise series, a mixture of chaotic series and colored noise generated from autoregressive process of order one is studied (results not shown). The same threshold value of $1/49$ beyond which the chaotic component can not be determined from a mixture of chaotic-signal and colored noise is observed.

The separation of the nonlinear deterministic signal from the random component is now considered. The separation is obtained through the use of predictive modeling strategies that have demonstrated value for chaotic systems. Several assumptions are made for this purpose. First, it is assumed that the functional form encapsulated by the trained nonlinear predictive models can be utilized to model the deterministic component of the time series. Thus, a decomposition of the series into a deterministic and a random component is realized. Second, it is assumed that the chaotic-signal is contained within the deterministic component and adequately modeled by the predictive models. Thus, if a functional form $\hat{\mathbf{f}}_{\mathbf{NN}}$ modeled by the ANN exists, where $\hat{z}_{t+T} = \hat{\mathbf{f}}_{\mathbf{NN}}(z_t, z_{t-\tau}, .., z_{t-(m-1)\tau}) + \epsilon$, the ANN can be trained using the entire data set to find $\hat{\mathbf{f}}_{\mathbf{NN}}$. Once trained, the estimate of $\hat{z}_t$ provided by the ANN is assumed to represent the deterministic component, and the residual is assumed to represent noise. This assumption holds as long as the functional form modeled by the predictive model is valid. The second assumption implies that the chaotic-signal, if any, is contained within the deterministic signal isolated using the predictive modeling-based decomposition strategy. It is noted, however, that this approach does not distinguish between chaotic and non-chaotic determinism. Once

the deterministic component has been isolated, the presence of chaos can be determined using the correlation dimension method. For the simulated data, a mixture of chaotic-signal and white noise is considered. Thus, the deterministic signal that can be isolated is assumed to be chaotic, provided the signal has been adequately modeled. The separated components can also be compared with the original signals from which the simulations are generated. Table 14 shows the comparison when the PSR is used as the predictive model, while Table 15 shows the same for the ANN. The recovered noise is further tested for normality. A perfect recovery would be indicated by a passing of this test with a mean that is statistically indistinguishable from zero and a standard deviation indistinguishable from that of the original white noise. Once the deterministic component has been isolated, it is compared with the original chaotic series used for the simulation. The prediction skills are related to *MSE* and *CC* values obtained from the predictive models. The results presented in Table 14 and 15 indicate that for the simulated data, a fairly good degree of separation can be achieved. An embedding dimension of 10 is used at which the saturation value of the correlation exponent of the Lorenz time series is obtained. When these concepts are generalized to the real data, the nature of the deterministic component is not known *a priori* and needs to be verified using tools like correlation exponent methodologies. A mixture of chaotic series and white noise ($\sigma = 0.17$) is analyzed for multi-step ahead forecasting using the PSR with the embedding dimension of 10. Fig. 57 shows the CC vs. $Forecast\ lead\ time$ and MSE vs. $Forecast\ lead\ time$ plots. It is observed that CC decreases and MSE increases as the forecast lead time increases. Due to randomness, the accuracy of the prediction for a chaotic series mixed with white noise decreases with an increase in the forecast lead time.

Table 14. Separation of white noise from a mixture of Lorenz (X component) and white noise using the PSR with $m = 10$.

σ_{wn}	σ_{wn}(cal)	% error	μ (norm fit)	μci (norm fit)	σci (norm fit)	MSE	CC
0.075	0.0861	14.8	-0.0012	-0.0066 - 0.0041	0.0825 - 0.0901	0.0074	0.9999
0.1	0.1066	6.6	-0.0022	-0.0088 - 0.0045	0.1021 - 0.1115	0.0114	0.9999
0.2	0.2008	0.4	0.0	-0.0135 - 0.0115	0.1924 - 0.2101	0.0403	0.9997
0.33	0.3502	6.12	-0.0011	-0.0229 - 0.0208	0.3355 - 0.3664	0.1226	0.9991

Table 15. Separation of white noise from a mixture of Lorenz (X component) and white noise using the ANN with $m = 10$.

σ_{wn}	σ_{wn}(cal)	% error	μ (norm fit)	μci (norm fit)	σci (norm fit)	MSE	CC
0.075	0.0871	16.1	0.0	-0.0054 - 0.0054	0.0834 - 0.0911	0.0076	0.9999
0.1	0.1129	12.9	0.0	-0.007 - 0.007	0.1081 - 0.1181	0.0138	0.9999
0.2	0.2238	11.9	0.0	-0.014 - 0.014	0.2144 - 0.2341	0.05	0.9996
0.33	0.3541	7.3	0.0	-0.0221 - 0.0221	0.3392 - 0.3704	0.1391	0.9991

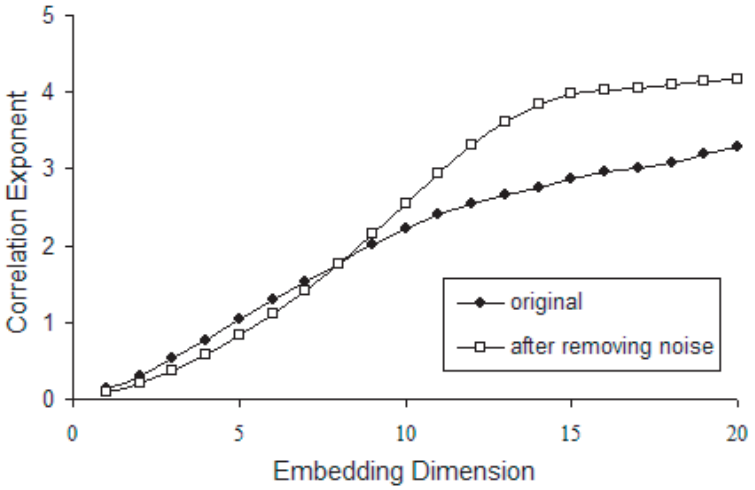

Figure 59. Correlation exponent vs. embedding dimension for monthly streamflow series at the Arkansas river.

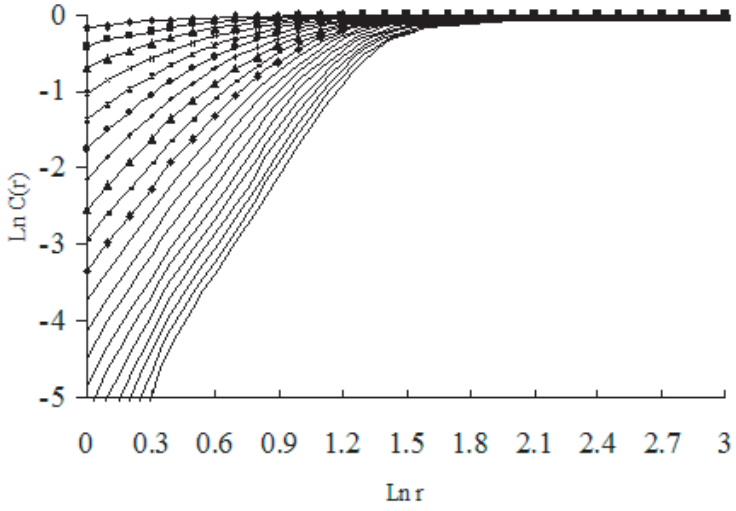

Figure 60. $LnC(r)$ vs. Lnr plot for the series, after removing noise, observed at the Arkansas river. The curves are shown in ascending order of embedding dimension, m = 1, 2,.., 20 from top to bottom.

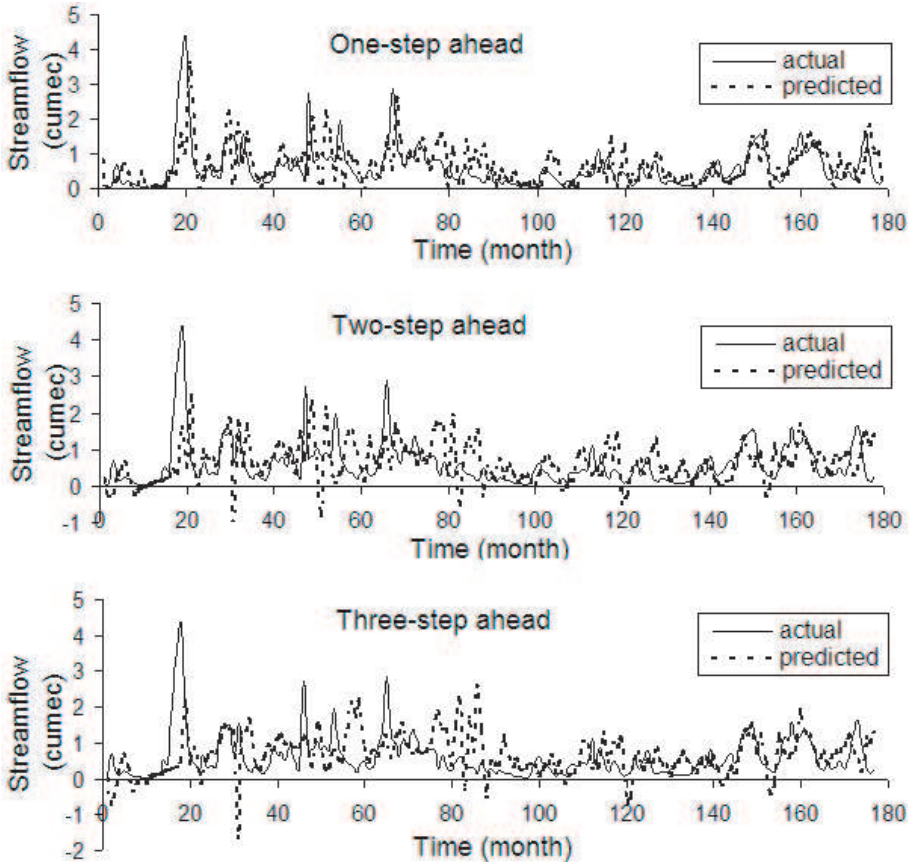

Figure 61. Multistep ahead predictions for the Arkansas river streamflow data. Top: one-step ahead predictions. Middle: two-step ahead predictions. Bottom: three-step ahead predictions.

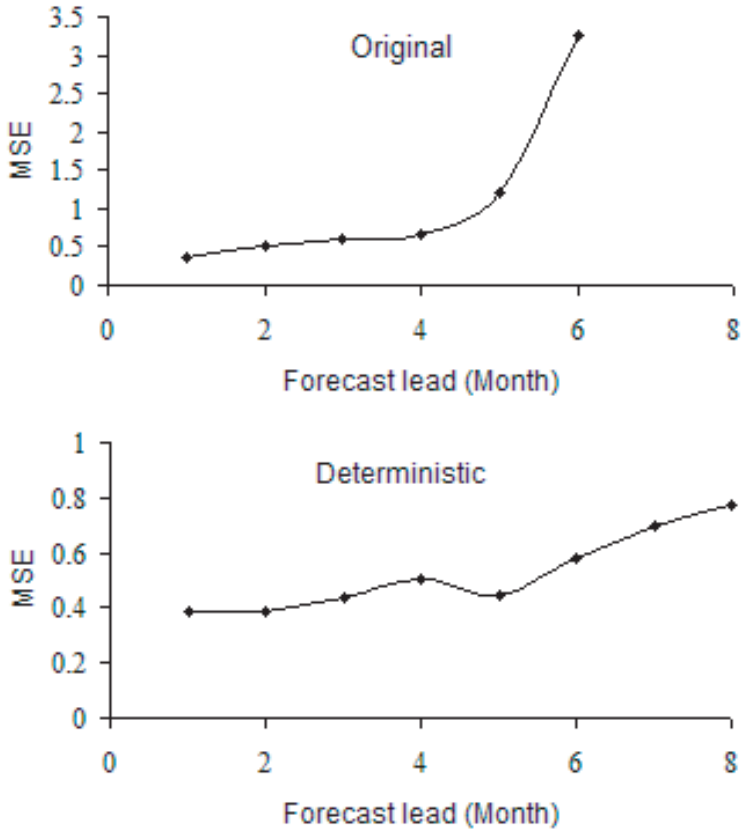

Figure 62. Monthly streamflow data at the Arkansas river: Variation of MSE with forecast lead time for the original and deterministic data. The deterministic data is obtained after removing noise from original data.

6.5 Analysis with hydrologic time series

6.5.1 Arkansas river

The monthly streamflow time series at the Arkansas river is shown in Fig. 50 and is analyzed using the correlation dimension method to detect the presence (or absence) of low-dimensional chaos in the system. The methodology used for the analysis is described in section 2.1.1. A moving window of size 3 units is used to compute the slope. The number of $Ln\ r$ values for which the slope is available and constant is limited for this data. The slope at a middle range (around $Ln\ r = 0.8$) is used as the correlation exponent. Fig. 59 shows the $correlation\ exponent$ vs. m plot for the original monthly streamflow series which does not saturate based either on an eye estimation or the objective techniques described in section 2.1.1. This indicates that either the system is stochastic or the chaotic component, if present, is dominated by noise or seasonality. The time series is analyzed using the PSR with $m = 10$ to isolate the deterministic component from the original series. The methodology is similar to that used for simulated data, and is described in detail in section 4.2. The mean and standard deviation of the non-deterministic component, i.e. noise, separated using the PSR are -0.0097 and 0.7159, respectively. The mean and standard deviation of the original series and the series after the removal of noise, i.e. deterministic series, are $\mu = 0.6529; \sigma = 0.6833$ and $\mu = 0.5908; \sigma = 0.6564$, respectively. In the present example, the ratio of σ_n (0.7159) and σ_c (0.6529) is 1.09 which is greater than the threshold value (i.e, 0.02) obtained for the simulated data in section 4.2.2. An attempt to fit a periodic function to the deterministic series to isolate seasonal behavior is not made since no dominant frequency in the series is observed from frequency domain analysis. The deterministic series is analyzed using the correlation dimension method with the same embedding dimension, i.e. 10, and window size, i.e. 3, to examine a low-dimensional chaos in the system. Fig. 60 shows the $Ln\ C(r)$ vs. $Ln\ r$ plot for the series. The $correlation\ exponent$ vs. m plot in Fig. 59 for the deterministic series shows that the correlation exponent saturates after an embedding dimension of 15 based on the objective techniques. The saturation value of the correlation exponent is 3.97. The number of variables required to model the streamflow dynamical system would be 4. The low value of correlation dimension suggests the possible presence of low-dimensional chaos in the streamflow dynamics. Fig. 61 shows one-step, two-step, and three-step ahead predictions for the original monthly streamflow series at the Arkansas river. Fig. 62 shows MSE vs. $Forecast\ lead\ time$ plots for the original and deterministic series. The MSE values at $Forecast\ lead\ time$ of 7 and 8 are 27.22 and 315.27, respectively. For the original series, MSE values are high and increase as the forecast lead time increases indicating the presence of non-deterministic component confirming the results obtained earlier

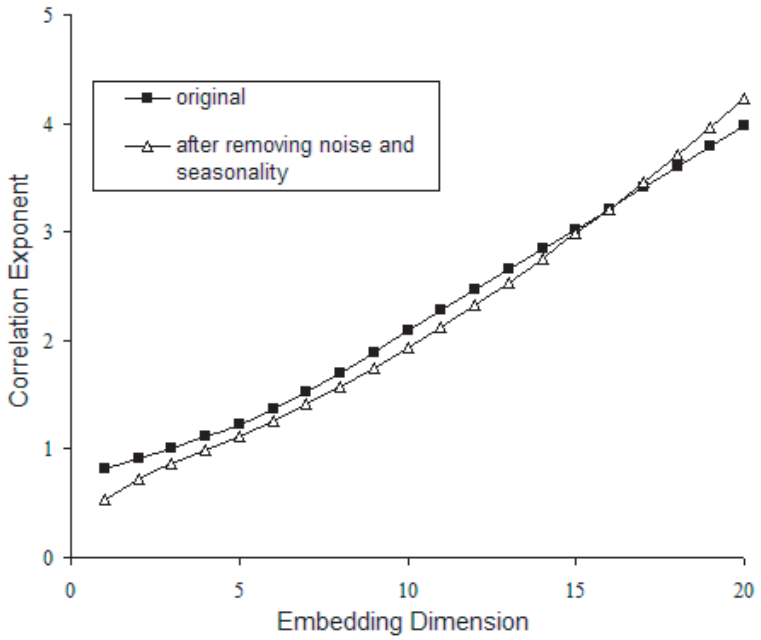

Figure 63. Correlation exponent vs. embedding dimension plot for daily streamflow series at the Colorado river.

using the PSR. This, in turn, decreases the prediction accuracy with an increase in the lead time. For the deterministic series, the prediction accuracy is significantly improved, i.e. the MSE values are very low as compared with the original series values, indicating that the deterministic series is indeed chaotic. This confirms the results from the correlation dimension method.

6.5.2 Colorado river

The daily streamflow series observed at the Colorado river below Parker dam is shown in Fig. 51. This series is analyzed for low-dimensional chaotic behavior using the correlation dimension method. The slope is calculated by considering a moving window of size 3 units. The correlation exponent is the slope value at $Ln\, r = -2.2$. The *correlation exponent* vs. m plot for the original series is shown in Fig. 63. Based on an eye estimation and the objective techniques described in section 2.1.1, it is seen that the correlation exponent does not reach a saturation value. This indicates that either the system is stochastic or it contains some dominant noise/seasonality component. To separate the non-deterministic component, the original series with $\mu = 0.1672$ and $\sigma = 0.0594$ is analyzed using the PSR with $m = 10$. The mean and the standard deviation of the non-deterministic component, i.e. noise, separated using the PSR are 0.0 and 0.0099, respectively. The statistics for the deterministic component that might contain chaos and seasonality are given as $\mu = 0.1669$

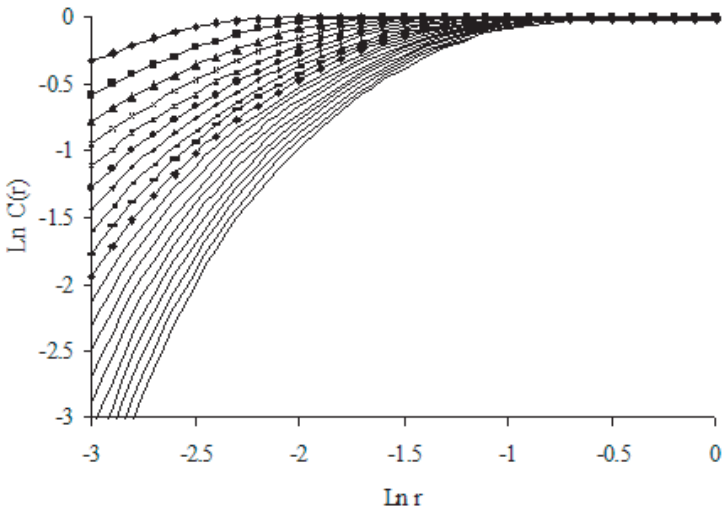

Figure 64. $LnC(r)$ vs. Lnr for daily streamflow series, after removing noise and seasonality, at the Colorado river. The curves are shown in ascending order of embedding dimension, m = 1, 2,.., 20 from top to bottom.

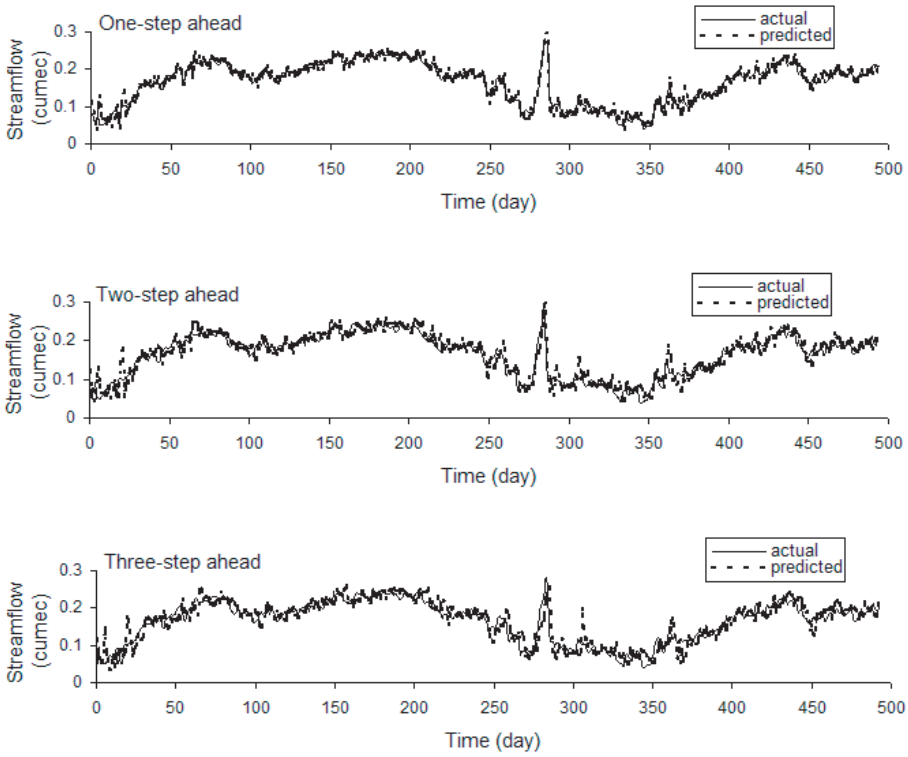

Figure 65. Multistep ahead predictions for the Colorado river streamflow data. Top: one-step ahead predictions. Middle: two-step ahead predictions. Bottom: three-step ahead predictions.

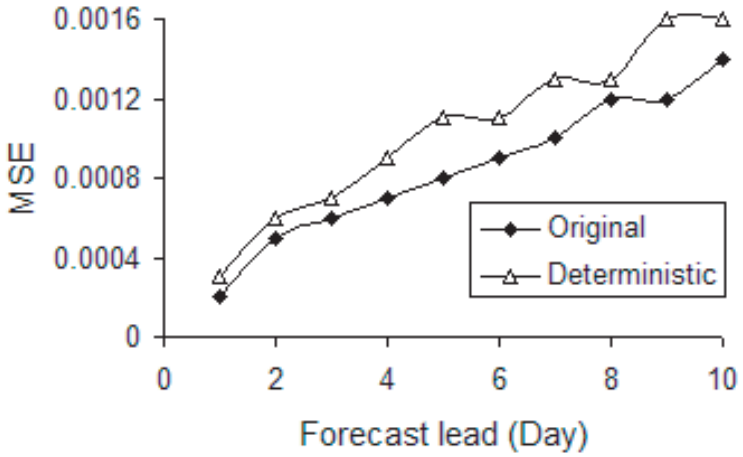

Figure 66. Daily streamflow data at the Colorado river: Variation of MSE with forecast lead time for the original and deterministic data. The deterministic data is obtained after separating white noise and seasonality from the original data.

and $\sigma = 0.0593$. The ratio of σ_n (0.0099) and σ_c (0.0593) is 0.167 which is greater than the threshold value of 0.02, obtained for the simulated data in section 4.2.2. The series, after the removal of noise, is analyzed using the frequency domain method to determine the dominant frequency. No peak is observed in the periodogram. However, based on an eye estimation a possible seasonal behavior in the data is observed. To isolate and remove the possible presence of seasonality, a periodic function is chosen to fit the data such that the MSE is minimized. This function is obtained using a trial and error procedure. The mean and the standard deviation of the series after the removal of noise and seasonality, i.e. deterministic series, according to the methods described earlier are 0.0469 and 0.0328, respectively. The deterministic series is analyzed using the correlation dimension method with the same embedding dimension, i.e. 10, and window size, i.e. 3. Fig. 64 shows the $Ln\ C(r)$ vs. $Ln\ r$ plot for the deterministic series. The $correlation\ exponent$ vs. m curve shown in Fig. 63 for the deterministic series does not saturate based on an eye estimation and the objective techniques indicating that the system is stochastic. Fig. 65 shows one-step, two-step, and three-step ahead predictions for the original daily series at the Colorado river. The MSE vs. $Forecast\ lead\ time$ plots for the original and deterministic series are shown in Fig. 66. The prediction accuracy is high, i.e. MSE values are low, due to the presence of seasonality in the data. For the original series, the MSE increases with an increase in the forecast lead time. The prediction accuracy does not change significantly for the deterministic series, i.e. MSE values are very close to the original series values, indicating the absence of chaos in the deterministic series. This indicates that the system is stochastic confirming the results from

the correlation dimension method. The streamflow measurement gauge is located at a distance of 100 feet downstream of the dam ([145]). In addition, the flow at the dam is regulated based on requirements ([146]) . These, in turn, may depend on various factors on a day-to-day basis leading to significant randomness in the observations. The measured flow at the dam, thus, reflects a variety of human controls implying that natural variability no longer is the dominant factor. This is the most likely cause for the loss of nonlinear dynamical information in the time series.

6.6 Summary and conclusions

The strengths of the use of nonlinear dynamical tools in conjunction with statistical time and frequency domain methodologies were demonstrated in this work. This dissertation investigated the ability to detect and model chaos from finite hydrologic observations, especially when randomness and seasonality are present. The ability to detect and model nonlinear dynamical and chaotic components, from finite real-world observations, is likely to have significant implications for scientific understanding and predictive modeling in multiple disciplines. This research represents a step forward in these directions. The results of the present investigation demonstrated the presence of thresholds, expressed in terms of noise to chaotic-signal and seasonality to chaotic-signal ratios. The ability to detect chaos from observations depends on whether the chaotic component in the hydrologic time series is dominant enough to satisfy the thresholds. It was shown that the overall time series can be decomposed into the contributing random, seasonal and chaotic components. Time series was decomposed using nonlinear predictive modeling for separating the chaotic component, statistical methods for characterizing random data, and frequency domain approaches for isolating seasonality. This has direct implications for a scientific understanding of hydrologic phenomena and the dominant processes that may be present as well as in the development of predictive models. For example, it was shown that the chaotic component, once detected and isolated, can be better predicted in the short-term through nonlinear models like ANN and PSR. The insights obtained from simulated data were used to interpret the results of real streamflow data from the Arkansas and Colorado rivers. It was observed that a chaotic component can be detected, isolated and utilized for improved predictive modeling from finite hydrologic time series like the Arkansas data. However, it was seen that for certain hydrologic data like the Colorado river data, this may not be possible. This may be due to the absence of chaotic/nonlinear dynamical component in the data. If the data observed at the Colorado river does contain chaotic/nonlinear dynamic component, then it may be completely dominated by randomness introduced as a result of the stochastic mode of dam operation on a daily basis.

Acknowledgements

Most of this work was done when the second author was on a visiting faculty appointment at the University of South Florida. The authors would like to thank two anonymous reviewers for their helpful comments.

Chapter 7

Conclusions

The application of different tools, specifically nonlinear dependence, extremes, and chaos, on the real data indicates that these tools may be easily fit as different components in developing predictive models for hydrology and climate. The insights obtained from the real data analysis using these tools are described below separately.

Rigorous analysis of recently developed MI-estimation methods indicate that two MI-estimation methods, specifically KDE and KNN, outperform the other methods and estimation procedures in terms of their ability to capture the dependence structure including nonlinear dependence where present. We find that KNN is the best estimator for *very short* data with relatively low noise while KDE works better for *very short* data when the noise levels are higher. For *short* data, KNN is the best choice for capturing the nonlinear dependence across all noise levels except when the data are generated from chaotic dynamics, where KDE is a better choice at higher noise levels. We surmise that the relative performance of KDE and KNN with respect to various noise levels is a consequence of the bias-variance tradeoff. The bias in the KDE estimates dominates the variance of the estimates for low noise-to-signal ratios. The KNN performs relatively better for low noise levels since its bias and variance are lower than that from KDE. However, the converse is true for high noise-to-signal ratios, and hence the KDE performs relatively better. For high noise, the variance dominates because of the noise in the data but the variance associated with $k = 3$ for KNN increases dramatically.

The application of MI-based nonlinear dependence on the real data suggests that there exists a nonlinear extrabasinal connection between ENSO and river flows in the topics and subtropics. This study also shows an appreciable increase of 20-70% in the variation of annual river flows linked to ENSO using nonlinear relationship measure as compared to linear measures. Hence, these results indicate additional predictability in the ENSO-streamflow extrabasinal connection when MI-based approaches are used, as compared to linear approaches used by researchers till date. The additional dependence captured by the MI-based nonlinear CCs may be useful for developing more accurate and longer streamflow models. Although ENSO has a direct influence on rainfall anomalies over the tropical and subtropical regions, only a portion of the variation in the annual flow of rivers located in these regions is associated with ENSO events. This may be due to

the complex relationship between rainfall and runoff, which, in turn, depends on surface hydrological and ocean-atmosphere-land interaction processes as well as noisy and potentially incomplete or corrupted data.

Analysis of precipitation extremes in South America indicate that the areas of interest based on an increasing PEVI from 1965-2004, are the Amazon basin, the Brazilian Highlands, Venezuela, Uruguay, Paraguay, and some of the highly populated cities in South America, specifically Rio De Janeiro, São Paulo, Caracas, Asuncion, and Cayenne. Some parts of east Brazil, few parts of the Brazilian highlands, São Paulo, and Cayenne also experience increasing number of consecutive 2- and 3-days extremes. Water resources engineers and planners, disaster management agencies, and policy makers need to pay special attention to the regions with increasing trends in the PEVI and consecutive 2- and 3-days daily extremes, especially when these regions overlap with densely populated areas, while planning for infrastructure development and disaster management.

Analysis of chaos indicate that the chaotic component, once detected and isolated, can be better predicted in the short-term through nonlinear models like ANN and PSR. The insights obtained from simulated data are used to interpret the results of real streamflow data from the Arkansas and Colorado rivers. It is observed that a chaotic component can be detected, isolated and utilized for improved predictive modeling from finite hydrologic time series like the Arkansas data. However, for certain hydrologic data like the Colorado river data, this may not be possible. This may be due to the absence of chaotic/nonlinear dynamical component in the data. If the data observed at the Colorado river does contain chaotic/nonlinear dynamic component, then it may be completely dominated by randomness introduced as a result of the stochastic mode of dam operation on a daily basis.

The above results also indicate the possibility of inter-connecting these tools for the purpose of developing predictive models. Since the geophysical data are generally noisy and MI-estimation methods are sensitive to noise, the noisy component in the data can be reduced, before applying MI-estimation methods for capturing nonlinear dependence, using ANN or PSR provided there is a presence of chaos in the data. This procedure may enhance the capability of MI-estimation methods for capturing nonlinear dependence. Nonlinear dependence between variables can help in identifying critical prediction variables out of many predictors and also enhance scientific understanding of relationships between different geophysical processes. The identification of relationships between variables may help in identifying additional variables, also called covariates, for extreme value distributions in order to improve extreme value models. There is a caveat that the use of covariates, having greater dependence, in extreme value distributions may not necessarily improve extreme value models.

References

[1] M. Barahona and C.-S. Poon. Detection of nonlinear dynamics in short, noisy time series. *Nature*, 381:215–217, 1996.

[2] Y. Moon, B. Rajagopalan, and U. Lall. Estimation of mutual information using kernel density estimators. *Phys. Rev. E*, 52(3):2318–2321, 1995.

[3] G. A. Darbellay and I. Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Trans. Inform. Theory*, 45(4):1315–1321, 1999.

[4] N. Kwak and C.-H. Choi. Input feature selection by mutual information based on Parzen window. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(12):1667–1671, 2002.

[5] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, 2004.

[6] M. M. V. Hulle. Edgeworth approximation of multivariate differential entropy. *Neural Computation*, 17:1903–1910, 2005.

[7] Q. Wang, Y. Shen, and J. Q. Zhang. A nonlinear correlation measure for multivariable data set. *Physica D*, 200:287–295, 2005.

[8] C. J. Cellucci, A. M. Albano, and P. E. Rapp. Statistical validation of mutual information calculations: Comparisons of alternative numerical algorithms. *Phys. Rev. E*, 71:066208, 2005.

[9] A. M. Fraser and H. L. Swinney. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A*, 33(2):1134–1140, 1986.

[10] S. Khan, S. Bandyopadhyay, A. R. Ganguly, S. Saigal, D. J. Erickson III, V. Protopopescu, and G. Ostrouchov. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Phys. Rev. E*, Accepted, 2007.

[11] M. P. Hoerling, A. Kumar, and M. Zhong. El Niño, La Niña, and the nonlinearity of their teleconnections. *J. Clim.*, 10:1769–1786, 1997.

[12] Y. H. Jin, A. Kawamura, K. Jinno, and R. Berndtsson. Nonlinear multivariable analysis of SOI and local precipitation and temperature. *Nonlinear Processes in Geophysics*, 12:67–74, 2005.

[13] E. A. B. Eltahir. El Niño and the natural variability in the flow of the Nile River. *Water Resour. Res.*, 32(1):131–137, 1996.

[14] K. N. Amarasekera, R. F. Lee, E. R. Williams, and E. A. B. Eltahir. Enso and the natural variability in the flow of tropical rivers. *J. Hydrology*, 200:24–39, 1997.

[15] D. W. Whitaker, S. A. Wasimi, and S. Islam. The El Niño-Southern Oscillation and long-range forecasting of flows in the Ganges. *Int. J. Climatol.*, 21:77–87, 2001.

[16] F. Anctil and P. Coulibaly. Wavelet analysis of the interannual variability in southern Québec streamflow. *J. Clim.*, 17:163–173, 2004.

[17] J. E. Richey, C. Nobre, and C. Deser. Amazon river discharge and climate variability: 1903-1985. *Science*, 246:101–103, 1989.

[18] G. Wang and E.A.B. Eltahir. Use of ENSO information in medium- and long-range forecasting of the Nile floods. *J. Clim.*, 12:1726–1737, 1999.

[19] E. Tziperman, L. Stone, M. A. Cane, and S. Zebiak. El Niño chaos: Overlapping of resonances between the seasonal cycle and the Pacific ocean-atmosphere oscillator. *Science*, 264:72–74, 1994.

[20] S.-I. An and F.-F. Jin. Nonlinearity and symmetry of ENSO. *J. Clim.*, 17:2399–2412, 2004.

[21] S. Khan, A. R. Ganguly, and S. Saigal. Detection and predictive modeling of chaos in finite hydrological time series. *Nonlinear Processes in Geophysics*, 12:41–53, 2005.

[22] T. R. Karl, R. W. Knight, and N. Plummer. Trends in high-frequency climate variability in the twentieth century. *Nature*, 377:217–220, 1995.

[23] P. Y. Groisman, T. R. Karl, D. R. Easterling, R. W. Knight, P. F. Jamason, K. J. Hennessy, R. Suppiah, C. M. Page, J. Wibig, K. Fortuniak, V. N. Razuvaev, A. Douglas, E. Forland, and P.-M. Zhai. Changes in the probability of heavy precipitation important indicators of climatic change. *Climatic Change*, 42:243–283, 1999.

[24] B. N. Goswami, V. Venugopal, D. Sengupta, M. S. Madhusoodanan, and P. K. Xavier. Increasing trend of extreme rain events over India in a warming environment. *Science*, 314:1442–1445, 2006.

[25] M. J. Manton, P. M. Della-Marta, M. R. Haylock, K. J. Hennessy, N. Nicholls, L. E. Chambers, D. A. Collins, G. Daw, A. Finet, D. Gunawan, K. Inape, H. Isobe, T. S. Kestin, P. Lefale, C. H. Leyu, T. Lwin, L. Maitrepierre, N. Ouprasitwong, C. M. Page, J. Pahalad, N. Plummer, M. J. Salinger, R. Suppiah, V. L. Tran, B. Trewin, I. Tibig, and D. Yee. Trends in extreme daily rainfall and temperature in Southeast Asia and the South Pacific: 1961-1998. *Int. J. Climatol.*, 21:269–284, 2001.

[26] R. Suppiah and K. J. Hennessy. Trends in total rainfall, heavy rainfall events, and number of dry events in Australia. *Int. J. Climatol.*, 18(10):1141–1164, 1998.

[27] M. Haylock and C. Goodess. Interannual variability of European extreme winter rainfall and links with mean large-scale circulation. *Int. J. CLimatol.*, 24:759–776, 2004.

[28] T. C. Peterson, M. A. Taylor, R. Demeritte, D. L. Duncombe, S. Burton, F. Thompson, A. Porter, M. Mercedes, E. Villegas, R. S. Fils, A. K. Tank, A. Martis, R. Warner, A. Joyette, W. Mills, L. Alexander, and B. Gleason. Recent changes in climate extremes in the Caribbean region. *J. Geophys. Res.*, D21:4601, doi:10.1029/2002JD002251, 2002.

[29] M. Brunetti, M. Maugeri, T. Nanni, and A. Navarra. Droughts and extreme events in regional daily Italian precipitation series. *Int. J. CLimatol.*, 22:543–558, 2002.

[30] T. Cavazos. Using self-organizing maps to investigate extreme climate events: An application to wintertime precipitation in the Balkans. *J. Climate*, 13(10):1718–1732, 2000.

[31] T. Iwashima and R. Yamamoto. A statistical analysis of the extremes events: Long-term trend of heavy daily precipitation. *J. Meteorol. Soc. Japan*, 71:637–640, 1993.

[32] C. Hellstrom and B. A. Malmgren. Spatial analysis of extreme precipitation in Sweden 1961-2000. *AMBIO: A Journal of the Numan Environment*, 33(4):187–192, 2004.

[33] L. M. V. Carvalho, Jones. C., and B. Liebmann. Extreme precipitation events in southeastern South America and large-scale convective patterns in the South Atlantic convergence zone. *J. Clim.*, 15:2377–2394, 2002.

[34] B. Liebmann, C. Jones, and L. M. V. Carvalho. Interannual variability of daily extreme precipitation events in the state of São Paulo, Brazil. *J. Climate*, 14:208–218, 2001.

[35] G. Kuhn. *On Dependence and Extremes*. PhD thesis, Munich University of Technology, 2006.

[36] G. Kuhn, S. Khan, A. R. Ganguly, and M. L. Branstetter. Geospatial-temporal dependence among weekly precipitation extremes with applications to observations and climate model simulations in South America. *Advances in Water Resources*, doi: 10.1016/j.advwatres.2007.05.006, 2007.

[37] A. F. Jenkinson. The frequency distribution of the annual maxima (or minima) values of meteorological elements. *Q. J. Meteorol. Soc.*, 81:158–171, 1955.

[38] E. J. Gumbel. *Statistics of extremes*. Columbia University Press, New York, 1958.

[39] R. W. Katz, M. B. Parlange, and P. Naveau. Statistics of extremes in hydrology. *Advances in Water Resources*, 25:1287–1304, 2002.

[40] S. Nadarajah. Extremes of daily rainfall in west cenral Florida. *Climate Change*, 69:325–342, 2005.

[41] R. W. Katz, G. S. Brush, and M. B. Parlange. Statistics of extremes: Modeling ecological disturbances. *Ecology*, 86(5):1124–1134, 2005.

[42] P. Todorovic and E. Zelenhasic. A stochastic model for flood analysis. *Water Resour. Res.*, 6:1641–1648, 1970.

[43] J. Pickands. Statistical inference using extreme order statistics. *Annals of Statistics*, 3:119–131, 1975.

[44] Y. Li, W. Cai, and E. P. Campbell. Statistical modeling of extreme rainfall in Southwest Western Australia. *J. Climate*, 18:852–863, 2005.

[45] P. S. Wilson and R. Toumi. A fundamental probability distribution of heavy rainfall. *Geophys. Res. Lett.*, 32(14):L14812,10.1029/2005GL022465, 2005.

[46] S. Khan, G. Kuhn, A. R. Ganguly, and D. J. Erickson III. Spatio-temporal variability of daily and weekly precipitation extremes in South America. *Water Resour. Res.*, In revision, 2007.

[47] S. D. Gaines and M. W. Denny. The largest, smallest, highest, lowest, longest, and shortest: extremes in ecology. *Ecology*, 74:1677–1692, 1993.

[48] I. Rodriguez-Iturbe, F. B. De Power, M. B. Sharifi, and K. P. Georgakakos. Chaos in rainfall. *Water Resour. Res.*, 25(7):1667–1675, 1989.

[49] S. Islam, R. L. Bras, and I. Rodriguez-Iturbe. A possible explanation for low correlation dimension estimates for the atmosphere. *J. Appl. Meteorol.*, 32:203–208, 1993.

[50] A. W. Jayawardena and F. Lai. Analysis and prediction of chaos in rainfall and streamflow time series. *J. Hydrol.*, 153:23–52, 1994.

[51] J. Stehlik. Deterministic chaos in runoff series. *J. Hydrol. Hydromech*, 47:271–287, 1999.

[52] B. Sivakumar. Chaos theory in hydrology: important issues and interpretations. *J. Hydrol.*, 227(1-4):1–20, 2000.

[53] B. Sivakumar. Forecasting monthly streamflow dynamics in the western united states: A nonlinear dynamical approach. *Environmental Modeling and Software*, 18(8/9):721–728, 2003.

[54] B. Sivakumar. Chaos theory in geophysics: Past, present and future. *Chaos, Solitons and Fractals*, 19(2):441–462, 2004.

[55] B. Sivakumar, R. Berndtsson, J. Olsson, K. Jinno, and A. Kawamura. Dynamics of monthly rainfall-runoff process at the gota basin: A search for chaos. *Hydrology and Earth System Sciences*, 4(3):407–417, 2000.

[56] B. Sivakumar, R. Berndtsson, J. Olsson, and K. Jinno. Evidence of chaos in rainfall-runoff process. *Hydrological Sciences Journal*, 46(1):131–146, 2001.

[57] B. Sivakumar, R. Berndtsson, and M. Persson. Monthly runoff prediction using phase-space reconstruction. *Hydrological Sciences Journal*, 46(3):377–387, 2001.

[58] B. Sivakumar, R. Berndtsson, J. Olsson, and K. Jinno. Discussion on analysis of cross-correlated chaotic streamflows by elshorbagy et al. *Hydrological Sciences Journal*, 47(3):523–527, 2002.

[59] B. Sivakumar, M. Persson, R. Berndtsson, and C. B. Uvo. Is correlation dimension a reliable indicator of low-dimensional chaos in short hydrological time series? *Water Resources Research*, 38(2), 10.1029/2001WR000333:31–38, 2002.

[60] A. Elshorbagy. Noise reduction approach in chaotic hydrologic time series revisited. *Canadian Water Resources Journal*, 26(4):537–550, 2001.

[61] A. Elshorbagy, U.S. Panu, and S.P. Simonovic. Analysis of cross-correlated chaotic streamflows. *Hydrological Sciences Journal*, 46(5):781–794, 2001.

[62] A. Elshorbagy, S. P. Simonovic, and U. S. Panu. Estimation of missing streamflow data using principles of chaos theory. *Journal of Hydrology*, 255(1-4):123–133, 2002.

[63] A. Elshorbagy, S. P. Simonovic, and U. S. Panu. Noise reduction in chaotic hydrologic time series: Facts and doubts. *Journal of Hydrology*, 256(3-4):147–165, 2002.

[64] M. N. Islam and B. Sivakumar. Characterization and prediction of runoff dynamics: a nonlinear dynamical view. *Adv. Water Resour.*, 25:179–190, 2002.

[65] D. Schertzer, Y. Tchiginskaya, S. Lovejoy, P. Hubert, H. Bendjoudi, and M. Larchevque. Which chaos in the rainfall-runoff process? *Hydrological Sciences Journal*, 47(1):139–149, 2002.

[66] B. Sivakumar, R. Berndtsson, J. Olsson, and K. Jinno. Reply to which chaos in the rainfall-runoff process? by schertzer et al. *Hydrological Sciences Journal*, 47(1):149–158, 2002.

[67] K. Fraedrich. Estimating the dimensions of weather and climate attractors. *J. Atmos. Sci.*, 43:419–432, 1986.

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

[69] S. Basu and E. Foufoula-Georgiou. Detection of nonlinearity and chaocity in time series using the transportation distance function. *Phy. Lett. A*, 301:413–423, 2002.

[70] G. Wang. A conceptual modeling study on biosphere-atmosphere interactions and its implications for physically based climate models. *Journal of Climate*, 17(13):2572–2583, 2004.

[71] R. L. Bras and I. RodriguesIturbe. Rainfall generation: A non-stationary time varying multidimensional model. *Water Resour. Res.*, 12:450–456, 1976.

[72] Y. Tessier, S. Lovejoy, P. Hubert, D. Schertzer, and S. Pecknold. Multifractal analysis and modeling of rainfall and river flows and scaling, causal transfer functions. *J. geophysical Res.*, 101(D21):26427–26440, 1996.

[73] E. Douglas and A. P. Barros. Probable maximum precipitation estimation using multifractals: Application in the eastern united states. *J. Hydrometeorology*, 4(6):1012–1024, 2003.

[74] R. J. Kuligowski and A. P. Barros. Localized precipitation forecasts from a numerical weather prediction model using artificial neural networks. *Weather and Forecasting*, 13(4):1194–1204, 1998.

[75] R. J. Kuligowski and A. P. Barros. Using artificial neural networks to estimate missing rainfall data. *J. of American Water Resources Association*, 34(6):1–11, 1998.

[76] G. Kim and A. P. Barros. Quantitative flood forecasting using multisensor data and neural networks. *J. Hydrology*, 246:45–62, 2001.

[77] A. R. Ganguly and R. L. Bras. Distributed quantitative precipitation forecasting combining information from radar and numerical weather prediction model outputs. *J. of Hydrometeorology, American Meteorological Society*, 4(6):1168–1180, 2003.

[78] J. Theiler and P. E. Rapp. Re-examination of the evidence for low-dimensional, nonlinear structure in the human electroencephalogram. *Electroencephalogr. Clin. Neurophysiol.*, 98(3):213–222, 1996.

[79] K. Lehnertz and C. E. Elger. Can epileptic seizures be predicted? Evidence from nonlinear time series analysis of brain electrical activity. *Phys. Rev. Lett.*, 80(22):5019–5022, 1998.

[80] D. A. Smirnov and B. P. Bezruchko. Estimation of interaction strength and direction from short and noisy time series. *Phys. Rev. E*, 68:046209, 2003.

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

[82] N. Nicolaou and S. J. Nasuto. Comment on "Performance of different synchronization measures in real data: A case study on electroencephalographic signals". *Phys. Rev. E*, 72:063901, 2005.

[83] R. Q. Quiroga, A. Kraskov, and P. Grassberger. Reply to "Comment on 'Performance of different synchronization measures in real data: A case study on electroencephalographic signals'". *Phys. Rev. E*, 72:063902, 2005.

[84] S. Khan, A. R. Ganguly, S. Bandyopadhyay, S. Saigal, D. J. Erickson III, V. Protopopescu, and G. Ostrouchov. Nonlinear statistics reveals stronger ties between enso and the tropical hydrological cycle. *Geophys. Res. Lett.*, 33:L24402, doi:10.1029/2006GL027941, 2006.

[85] B. Rajagopalan, U. Lall, and D.G. Tarboton. Evaluation of kernel density estimation methods for daily precipitation sampling. *Stochastic Environmental Research and Risk Assessment*, 11(6):523–547, doi: 10.1007/BF02428432, 1997.

[86] G. A. Darbellay. An estimator of the mutual information based on a criterion for independence. *Computational Statistics and Data Analysis*, 32:1–17, 1999.

[87] D. R. Brillinger. Some data analyses using mutual information. *Brazilian J. Probability and Statistics*, 18:163–183, 2004.

[88] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley, New York, 1991.

[89] R. Steur, J. Kurths, C. O. Daub, J. Weise, and J. Selbig. The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, 81(2):S231–S240, 2002.

[90] J. Xu, Z.-R. Liu, R. Liu, and Q.-F. Yang. Information transmission in human cerebral cortex. *Physica D*, 106:363–374, 1997.

[91] T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85(2):461–464, 2000.

[92] M. S. Pinsker. *Information and information stability of random variables and processes.* San Francisco: Holden-Day, 1964.

[93] H. Joe. Relative entropy measures of multivariate dependence. *J. American Statistical Association*, 84(405):157–164, 1989.

[94] C. Granger and J. Lin. Using the mutual information coefficients to identify lags in nonlinear models. *J. Time Series Analysis*, 15(4):371–384, 1994.

[95] S. J. Schiff, P. So, and T. Chang. Detecting dynamical interdependence and generalized synchrony through mutual prediction in a neural ensemble. *Physical Review E*, 54(6):6708–6724, 1996.

[96] H. K. M. Meeren, J. P. M. Pijn, E. L. J. M. Luijtelaar, A. M. L. Coenen, and F. H. L. Silva. Cortical focus drives widespread corticothalamic networks during spontaneous absence seizures in rats. *J. Neuroscience*, 22(4):1480–1495, 2002.

[97] B. W. Silverman. *Density estimation for statistics and data analysis.* Chapman and Hall/CRC, 1986.

[98] P. B. Wright. Homogenized long-period Southern Oscillation indices. *Int. J. Climatol.*, 9:33–54, 1989.

[99] M. C. Todd and R. Washington. Climate variability in central equatorial Africa: Influence from the Atlantic sector. *Geophys. Res. Lett.*, 31(23), 2004.

[100] G. Pizarro and U. Lall. El Niño and Floods in the US West: What can be expect? *EOS, Transactions of the AGU*, 83(32):349–352, 2002.

[101] M. Kendall and J. D. Gibbons. *Rank correlation methods.* A Charles Griffin, 5th Edition, 1990.

[102] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap.* Chapman and Hall/CRC, 1993.

[103] F. C. Curriero, J. A. Patz, J. B. Rose, and S. Lele. The association between extreme precipitation and waterborne disease outbreaks in the United Status, 1948-1994. *Am. J. Public Health*, 91(8):1194–1199, 2001.

[104] J. R. Stedinger and T. A. Cohn. Flood frequency analysis with historical and paleoflood information. *Water Resources*, 22:785–793, 1986.

[105] R. W. Katz. Stochastic modeling of huricane damage. *J. Appl. Meteorol.*, 41:754–762, 2002.

[106] B. G. Brown and R. W. Katz. Regional analysis of temperature extremes: spatial analog for climate change? *J. Climate*, 8:108–119, 1995.

[107] J. P. Palutikof, B. B. Brabson, D. H. Lister, and S. T. Adcock. A review of methods to calculate extreme wind speeds. *Meterological Applications*, 6:119–132, 1999.

[108] F. P. Schoenberg, R. Peng, and J. Woods. On the distribution of wildfire sizes. *Environmetrics*, 14:583–592, 2003.

[109] B. Liebmann and D. Allured. Daily precipitation grids for South America. *Bull. Amer. Meteor. Soc.*, 86(11):1567–1570, 2005.

[110] S. G. Coles. *An introduction to statistical modeling of extreme values.* Springer-Verlag, London, UK, 2001.

[111] D. R. Easterling, J. L. Evans, P. Ya. Groisman, T. R. Karl, K. E. Kunkel, and P. Ambenje. Observed variability and trends in extreme climate events. *Bull. Am. Met. Soc.*, 81:417–425, 2000.

[112] G. A. Meehl and C. Tebaldi. More intense, more frequent, and longer lasting heat waves in the 21st century. *Science*, 305:994–997, 2004.

[113] J. R. Michael. The Stabilized Probability Plot. *Biometrika*, 70(1):11–17, 1983.

[114] S. G. Coles. On goodness-of-fit tests for the two-parameter Weibull distribution derived from the stabilized probability plot. *Biometrika*, 76(3):593–598, 1989.

[115] A. C. Kimber. Tests for the Exponential, Weibull and Gumbel Distributions Based on the Stabilized Probability Plot. *Biometrika*, 72(3):661–663, 1985.

[116] J. Galambos. *The asymptotic theory of extreme order statistics*. Robert E. Krieger, Malabar, FLorida, USA, 1987.

[117] B. Rajagopalan and U. Lall. A k-nearest-neighbor simulator for daily precipitation and other weather variables. *Water Resour. Res.*, 35(10):3089–3101, 1999.

[118] D. Yates, S. Gangopadhyay, B. Rajagopalan, and K. Strzepek. A technique for generating regional climate scenarios using a nearest-neighbor algorithm. *Water Resour. Res.*, 39(7):1199, doi:10.1029/2002WR001769, 2003.

[119] H. D. Abarbanel. *Nonlinear Systems*. VCH Publishers, New York, 1994.

[120] E. N. Lorenz. Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, 26:636–646, 1969.

[121] K. Fraedrich. Estimating weather and climate predictability on attractors. *J. Atmos. Sci.*, 44:722–728, 1987.

[122] C. Essex, T. Lookman, and M. A. H. Nerenberg. The climate attractor on short time scales. *Nature*, 326:64–66, 1987.

[123] A. Hense. On the possible existence of a strange attractor for the southern oscillation. *Beitr Phy. Atmos.*, 60(1):34–47, 1987.

[124] B. P. Wilcox, M. S. Seyfried, and T. H. Matison. Searching for chaotic dynamics in snowmelt runoff. *Water Resour. Res.*, 27(6):1005–1010, 1991.

[125] E. N. Lorenz. *The Essence of Chaos (The Jessie and John Danz Lecture Series)*. University of Washington Press, 1996.

[126] A. Porporato and L. Ridolfi. Clues to the existence of deterministic chaos in river flow. *Int. J. Mod. Phys. B*, 10:1821–1862, 1996.

[127] Y. Almog, O. Oz, and S. Akselrod. Correlation dimension estimation: can this nonlinear description contribute to the characterization of blood pressure control in rats? *IEEE Transactions on Biomedical Engineering*, 46(5):535–537, 1990.

[128] D. Hsieh. Chaos and nonlinear dynamics: Applications to financial markets. *Journal of Finance*, 46:1839–1878, 1991.

[129] R. R. Trippi. *Chaos & Nonlinear Dynamics in the Financial Markets: Theory, Evidence and Applications*. Irwin Professional Publishing, 1995.

[130] A. L. Cornelis. Visualization of chaos for finance majors. In *2000 Finance Educators Conference: Finance Education in the New Millennium, Proceedings of the 2000 Annual Conference*, pages 187–226, Deakin University, Burwood, Victoria, Australia, 2000.

[131] G. S. Yim, J. W. Ryu, Y. J. Park, S. Rim, S. Y. Lee, W. H. Kye, and C. M. Kim. Chaotic behaviors of operational amplifiers. *Physical Review E*, 69, 2004.

[132] F. Takens. *Detecting strange attractors in turbulence, in Dynamical Systems and Turbulence, Lecture notes in Mathematics, 898, pp. 366-381.* Springer Verlag, New York, 1980.

[133] B. Sivakumar, S. Y. Liong, C. Y. Liaw, and K. K. Phoon. Evidence of chaotic behavior in singapore rainfall. *J. Am. Water Resour. Assoc.*, 34:301–310, 1998.

[134] R. C. Hilborn. *Chaos and Nonlinear Dynamics.* Oxford University Press, 2000.

[135] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D*, 9:189–208, 1983.

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

[137] A. R. Osborne and A. Provenzale. Finite correlation dimension for stochastic systems with power-law spectra. *Physica D*, 35:357–381, 1989.

[138] M. A. H. Nerenberg and C. Essex. Correlation dimension and systematic geometric effects. *Phys. Rev. Lett. A*, 42(12):7065–7074, 1990.

[139] ASCE. Task committee on applications of artificial neural networks in hydrology, ii, hydrologic applications. *J. Hydrol. Eng.*, 5(2):124–137, 2000.

[140] A. S. Weigend and N. A. Gershenfeld. *Time Series Prediction: Forecasting the Future and Understanding the Past.* Addison-Wesley, Reading, MA, 1994.

[141] G. Zhang, B. E. Patuwo, and M. Y. Hu. Forecasting with artificial neural networks: The state of the art. *Int. J. Forecasting*, 14:35–62, 1998.

[142] C. M. Bishop. *Neural Networks for pattern recognitions.* Oxford University Press, 1995.

[143] J. D. Farmer and J. J. Sidorowich. Predicting chaotic time series. *Phys. Rev. Lett.*, 59:845–848, 1987.

[144] D. Koutsoyiannis and D Pachakis. Deterministic chaos versus stochasticity in analysis and modeling of point rainfall series. *J. Geophys. Res.*, 101:26,441–26,451, 1996.

[145] USGS. *private communication.* 2004.

[146] BoR. *Bureau of Reclamation, private communication.* 2004.

About the Author

Shiraj Khan is a Ph.D. candidate in the Department of Civil and Environmental Engineering at the University of South Florida. He is expected to receive his Ph.D. degree on August 11, 2007. He received a Bachelor of Technology (B. Tech.) degree from Indian Institute of Technology (IIT), Roorkee in 2001. At IIT, he was the only student selected for *The President of India Gold Medal* for outstanding proficiency including character, conduct, excellence in academic performance, extra-curricular activities and social services. He also received *IIT Silver Medal* for getting the highest GPA in Civil Engineering. In 2005, he was selected by American Society of Engineers of Indian Origin for the *Kalpana Chawla-Ford Motor Company Award* in recognition of his high scholastic achievements in engineering. During his doctoral studies, he published six journal papers in the high quality journals and three book chapters, and presented several conference papers.