Detection, Classification and Analysis of Events in Turbulence Time Series

A thesis submitted for the degree of

Doctor of Philosophy

by

Yanfei Kang

School of Mathematical Sciences

Monash University

Australia

May 2014

Contents

Abstract

Time series is the major source of information to study characteristics of the atmospheric boundary layer (ABL), which is frequently dominated by various types of events embedded in the time series with different levels of noise. To understand physical dynamics of the atmospheric turbulence, the individual events need to be detected and studied about their physical and structural characteristics. In spite of the attention that has been given to studying events among the atmospheric science community, the detection of events still presents a challenge, and thus their characteristics and contributions to the ABL remain poorly understood. Besides the existence of high level noise in turbulence, the main difficulty is that many of the events that are responsible for the variability in the atmospheric turbulence time series are previously unknown, especially in the stable ABL.

This dissertation develops a new method for detecting and classifying structures from turbulence time series. The main idea of the method is in defining events as time-series subsequences that are significantly different from noise. This switches the focus of the event detection approach towards defining the characteristics of noise, which is in many situations an easier problem than defining a structure. For atmospheric time series, a natural characterization of the noise is red noise, which is a stationary AR(1) process. The proposed method consists of two steps. The first step of the method is event extraction based on noise tests. We perform a noise test on each subsequence extracted from the series using a sliding window.

All the subsequences recognized as noise are removed from further analysis, and the events are extracted from the remaining non-noise subsequences. This step does not assume particular geometries or amplitudes of the flow structures. In the second step, the detected structures are classified into groups with similar characteristics. This step groups large numbers of detected events such that it opens a pathway for the detailed study of their characteristics, and helps to gain understanding of events with previously unknown origin. In order to account for the underlying characteristics of the extracted events, a feature-based clustering method is used, which first summarizes each event with its global measures before performing clustering in the feature space. It yields substantially better results than clustering based on raw data of the events.

The developed R package **TED** is tested on artificial time series with different levels of complexity and real world atmospheric turbulence time series. The results on artificial data show that events used to generate the data can be exactly detected and clustered. The method is robust to high levels of noise, which is advantageous regarding very noisy turbulence time series. Application of the method to a well-known real world turbulence dataset demonstrates that the method successfully extracts realistic flow structures, which are in line with previous studies that have examined the underlying physical mechanisms of several isolated events on that dataset. From the application of the method to a more complicated turbulence dataset, about which no published results can be found regarding extraction of unknown events, the proposed method is able to detect and distinguish events with different dynamical characteristics even though the clustering step is only based on statistical measures of characteristics of events from time series.

List of publications

1. Kang Y. 2012. Real-time change detection in time series based on growing feature quantization. In: *Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–6, doi: 10.1109/IJCNN.2012.6252381

2. Kang Y, Belušić D, Smith-Miles K. 2014d. A note on the relationship between turbulent coherent structures and phase correlation. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **24**(2): 023114, doi: http://dx.doi.org/10.1063/1.4875260

3. Kang Y, Smith-Miles K, Belušić D. 2013. How to extract meaningful shapes from noisy time-series subsequences? In: *Proceedings of the 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE, pp. 65–72, doi: 10.1109/CIDM.2013.6597219

4. Kang Y, Belušić D, Smith-Miles K. 2014c. Detecting and classifying events in noisy time series. *Journal of the Atmospheric Sciences* **71**(3): 1090–1104, doi: 10.1175/JAS-D-13-0182.1

5. Kang Y, Belušić D, Smith-Miles K. 2014a. Classes of structures in the stable atmospheric boundary layer (*Submitted*). *Quarterly Journal of the Royal Meteorological Society*

PART A: General Declaration

Monash University

Declaration for thesis based or partially based on conjointly published or unpublished work

General Declaration

In accordance with Monash University Doctorate Regulation 17.2 Doctor of Philosophy and Research Master's regulations the following declarations are made:

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes 4 original papers published in peer reviewed journals and 1 submitted publication. The core theme of the thesis is event detection, classification and analysis in turbulence time series. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the candidate, working within the School of Mathematical Sciences, Monash University, under the supervision of Professor Kate Smith-Miles and Doctor Danijel Belusic.

The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.

In the case of Chapters 2-6 my contribution to the work involved the following:

Thesis chapter	Publication title	Publication status*	Nature and extent of candidate's contribution
2	Real-time change detection in time series based on growing feature quantisation	Published	Developed, established and verified the method Wrote programming codes and the article
3	A note on the relationship between turbulent coherent structures and phase correlation	In press	Developed, established and verified the method Wrote programming codes and the article
4	How to extract meaningful shapes from noisy time-series subsequences?	Published	Developed, established and verified the method Wrote programming codes and the article
5	Detecting and classifying events in noisy time series	Published	Developed, established and verified the method Wrote programming codes and the article
6	Classes of structures in the stable atmospheric boundary layer	Submitted	Developed, established and verified the method Wrote programming codes and the article

I have / have not (circle that which applies) renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Signed: ..... [redacted] .....

Date: **……21/05/2014……**

Acknowledgement

I would like to express my deepest gratitude to my PhD supervisors, Prof. Kate Smith-Miles and Dr. Danijel Belušić, for their patient guidance, enormous encouragement and constructive criticism throughout the past few years. Kate has taught me, both consciously and unconsciously, how good research is done. Her wisdom, knowledge and magnetic personality inspired and motivated me. I am also thankful for the excellent example she has provided as a successful female mathematician and professor. I am deeply grateful to Danijel, who has supported me not only by providing countless insightful discussions about the research, but also academically and emotionally through the tough road to finish this thesis. I have been extremely lucky to have a supervisor like him who cared so much about my work, and who responded to my questions and queries so promptly.

Also, thanks to Dr. Christian Rau, one of the nicest people imaginable, who co-supervised my first year research. My sincere appreciation also goes to Prof. Maozai Tian, my former Master's research advisor, whose collaboration with Christian provided me the opportunity to pursue a PhD in Monash University.

I am thankful to Prof. Rob Hyndman and Prof. Fima Klebaner, members of my PhD committee, for their helpful feedback and suggestions. A very special thanks to Prof. Larry Mahrt for his valuable comments on my research and also for providing the FLOSSII dataset. Dr. Jielun Sun is sincerely thanked for providing the CASES-99 data. Thanks also goes to Prof. Eamonn Keogh and his former PhD

student Abdullah Mueen for providing me the Matlab codes that have been used for methodological comparison in Chapter 4.

It is my pleasure to acknowledge all my current and previous colleagues in School of Mathematical Sciences, Monash University, for their company and friendship, especially Yie Yen Fan, Mark Bentley, Salem Alyami, Kareem Elgindy, Laura Villanova, Nur Insani, Carolina Segovia, Vivian Huang and Oscar Tian. Completing this work would have been more difficult without the friendly support provided by the administrative members of the school. I am indebted to them, especially Andrea Peres, Linda Mayer and Gertrude Nayak.

I must give my sincere thanks to my closest friends in this country, Qianqian Wu, Jenny Huang and Eric Zhou, for their company, caring and entertainment, in my life and studies. They were the persons I turned to in both good times and in bad. My greatest appreciation also goes to my best friends in China, Yan Sun, Yifang Han and Wei Han for their continuous mental support throughout my student life. They were also motivators who always asked my PhD research progress. Special thanks to Angela Qi in Canada for her endless encouragement and invaluable inspiration since we knew each other in 2006. I need to also thank Lingbing Feng, a PhD student in College of Business and Economics, Australia National University, for being a very good friend who frequently pushed me to learn more. He was also the first friend I wanted to talk with when I got stuck in my studies.

Finally, but not least, thanks goes to my family for all their continuous love and their supports in my decisions. My love and gratitude for them can hardly be expressed in words. I dedicate this thesis to my parents.

This research was supported by the China Scholarship Council.

List of Figures

List of Tables

Chapter 1
Introduction

Chapter 1

Introduction

1.1 Background

Turbulence is one of the most challenging and exciting areas of fluid mechanics. A complicated series of structures take place that eventually lead to the flow becoming turbulent. To understand how turbulence is generated and dissipated, its internal structure needs to be studied (Robinson, 1991). In a geophysical context, atmospheric turbulence is characterized by frequently occurring recognizable structures atop the more random background motions. These flow structures are usually responsible for the intrinsic phenomena of turbulence such as the transport of mass, heat and momentum, and exert significant effects on turbulent mixing and dissipative properties (Hussain, 1983; Bergström and Högström, 1989). They explain the rather organized and deterministic part of turbulence while the remaining component is represented by more random superimposed background motions (Hussain, 1983; Turner and Leclerc, 1994; Belušić and Mahrt, 2012; Campanharo *et al.*, 2008). Therefore, comprehensive studies of these flow structures provide a powerful tool to improve our knowledge of turbulence flows. Likewise, in order to better understand the turbulence in the atmospheric boundary layer

(ABL), the major question is how to isolate the structures from the background noise (Turner and Leclerc, 1994).

Over the last decades, despite a large body of knowledge that has been established in the field of turbulent structures, the detection of such structures in the ABL still presents a challenge. Firstly, high levels of noise are present in atmospheric turbulence, which significantly increases the fundamental complexity of turbulence phenomena. Secondly, many processes that characterize the variability in atmospheric turbulence time series are from previously unknown origins, especially in the stable ABL (Mahrt, 2011a). As a result, the complete picture of the flow structures and their dynamics remain poorly identified. However, these flow structures are responsible for the production of turbulence in the boundary layer and understanding them would significantly contribute to the general understanding of turbulent flows. They need to be detected from background fluctuations to get insights into the dynamical properties and behavior of the atmospheric turbulence flows in terms of elementary structures.

“Events" and “structures" are used interchangeably in this thesis since “event" is another name for a “structure". Structures are physical objects that we can be viewed as events in time series. Because events are not exactly defined or well known in literature, the focus of this thesis is to first detect events from time series, and then analyse their dynamics to better understand what they are.

1.2 Event Detection in Atmospheric Sciences

Since the existence of non-random structures in turbulence was recognized, many methods have been proposed to detect structures and study their nature. Some of the methods that have been used mostly for detection and location of events are summarized in this section. Each of them has unique criteria for definition of

events and therefore yields different results. The notation has been unified in the following description.

1.2.1 Variable Interval Time-Averaging (VITA)

The VITA method was firstly proposed by Blackwelder and Kaplan (1976) to detect the occurrence of events in turbulence. The basic idea is that the appearance of flow structures is supposed to result in larger fluctuation, which can be measured by variance of short-time averages. For a signal $f(s)$, its short-time variance $\widehat{\mathrm{var}}(t,T)$ at time t in the short-time-averaging interval T is defined as

$$\widehat{\mathrm{var}}(t,T) = \frac{1}{T}\int_{t-\frac{T}{2}}^{t+\frac{T}{2}} f^2(s)ds - \left(\frac{1}{T}\int_{t-\frac{T}{2}}^{t+\frac{T}{2}} f(s)ds\right)^2. \tag{1.2.1}$$

Blackwelder and Kaplan (1976) stated that the detection of events uses a threshold θ on the short-time variance $\widehat{\mathrm{var}}(t,T)$. When $\widehat{\mathrm{var}}(t,T)$ exceeds $\theta\sigma_f^2$, it indicates an event, where σ_f^2 is the variance of the signal $f(s)$. Accordingly, the event detecting function for VITA is

$$D_{VITA} = \begin{cases} 1, & \text{if } \widehat{\mathrm{var}}(t,T) \geq \theta\sigma_f^2, \\ 0, & \text{otherwise,} \end{cases} \tag{1.2.2}$$

1.2.2 Windowed Averaged Gradient (WAG)

The WAG method was introduced by Antonia and Fulachier (1989) as an alternative to VITA. In this method, the abrupt changes associated with flow events are detected by measuring the average gradient $\mathrm{Grad}(t,T)$ at time t over some time interval T:

$$\mathrm{Grad}(t,T) = \frac{1}{T}\left(\int_{t-\frac{T}{2}}^{t} f(s)ds - \int_{T}^{t+\frac{t}{2}} f(s)ds\right). \tag{1.2.3}$$

The definition of the event detecting function for WAG is similar to that in VITA:

$$D_{WAG} = \begin{cases} 1, & \text{if Grad}(t, T) \geq \theta\sigma_f, \\ 0, & \text{otherwise,} \end{cases} \tag{1.2.4}$$

where σ_f is the standard deviation of the signal $f(s)$.

The results of event detection using both VITA and WAG, to a large extent, depend on the choice of parameters: the threshold θ, which indicates the strength of detected structures, and the short-time interval T, which relates to the time scale. Usually the threshold is chosen by seeking a balance between better statistical accuracy, which is given by a smaller θ, and greater event significance given by a larger θ (Segalini and Alfredsson, 2012). Having chosen the two parameters, determination of the number of flow events is made according to the corresponding criterion.

1.2.3 Quadrant Analysis

Each of the three velocity components (streamwise u, cross-streamwise v, vertical w) can be separated into its mean value ($\overline{u}$, $\overline{v}$, $\overline{w}$) and the fluctuating component (u', v', w'). Quadrant analysis, firstly introduced by Wallace *et al.* (1972), aims to find large $|u'w'|$ values in the quadrant of the $u' - w'$ plane. The technique is based on the scatterplot of u' versus w', whose Cartesian axes define four quadrants:

- Quadrant 1 (Q1): $u' > 0$, $w' > 0$;
- Quadrant 2 (Q2): $u' < 0$, $w' > 0$;
- Quadrant 3 (Q3): $u' < 0$, $w' < 0$;
- Quadrant 4 (Q4): $u' > 0$, $w' < 0$.

The second quadrant Q2 and the fourth quadrant Q4 have been mostly used by researchers, which characterize “ejection-like" and “sweep-like" events, respectively (e.g., Wallace *et al.*, 1972; Rajagopalan and Antonia, 1982; Thomas and Foken, 2007a; Steiner *et al.*, 2011). This method has been used in the turbulence community to detect strong “sweep-like" or “ejection-like" events. The event detecting function is defined as:

$$D_Q = \begin{cases} 1, & \text{if } |u'w'| > \theta \text{ and } u'w' < 0 \\ 0, & \text{otherwise,} \end{cases} \tag{1.2.5}$$

The number and duration of events detected with quadrant analysis are sensitive to the threshold parameter θ. The number of events decreases quickly with larger θ (e.g., Bergström and Högström, 1989; Zhu *et al.*, 2007; Steiner *et al.*, 2011).

1.2.4 Wavelet Analysis

Since the introduction of the wavelet transform to turbulence research in Farge (1992) and the suitability of wavelets analysis for the detection of events in turbulence was verified in Collineau and Brunet (1993a,b), multiple wavelet methods have successfully identified events from turbulence data (e.g., Gilliam *et al.*, 2000; Chen and Hu, 2003; Thomas and Foken, 2005; Barthlott *et al.*, 2007; Segalini and Alfredsson, 2012). It has been rapidly accepted as an objective method to detect structures since it does not require specifications of parameters (e.g., specification of the threshold level is required in VITA, WAG and quadrant techniques). Wavelet analysis constructs a time-frequency representation of a signal using wavelet functions offering very good time and frequency localization, which in the detection of events provides information on both moment of occurrence (time) and the representative duration (frequency) of events. Essentially, wavelet transforms

reveal the relative contributions of different time scales to the overall fluctuation of the signal by decomposing a signal into scaled and translated wavelet functions. The continuous wavelet transform of a signal $x(t)$ can be represented as:

$$W_b(a) = \frac{1}{a}\int_{-\infty}^{+\infty} x(t)\psi\left(\frac{t-b}{a}\right)dt, \tag{1.2.6}$$

where $\psi(t)$ is the chosen mother wavelet function, a is the wavelet scale, b is position translation, and $W_b(a)$ represents the wavelet coefficients.

The choice of wavelet function, which determines the final waveform shape, is particularly important since the differences between different mother wavelet functions influence how the scaled signals and the wavelets are defined. Collineau and Brunet (1993a) compared four wavelet functions (Mexican-Hat, HAAR, RAMP and WAVE) for detection of jumps and concluded that the Mexican-Hat wavelet outperforms others on the tested time series. Since then, several studies have used the Mexican-Hat wavelet to detect structures in turbulence data (e.g., Chen *et al.*, 1997; Thomas and Foken, 2005; Barthlott *et al.*, 2007; Feigenwinter and Vogt, 2005).

The mean time scale of events can be found by finding the peaks of the global wavelet spectrum over all time scales (e.g., Collineau and Brunet, 1993a; Mahrt, 1991; Mahrt and Gibson, 1992). The zero-crossings of the wavelet coefficients at the scale of the maximum global wavelet spectrum provide information on the moments of event occurrence. The global wavelet spectrum (also known as wavelet variance), introduced by Collineau and Brunet (1993a), can be obtained by integrating the square modulus of the wavelet coefficients over all translations for each wavelet scale:

$$\overline{W(a)} = \frac{1}{a}\int_{-\infty}^{+\infty} |W_b(a)|^2 db. \tag{1.2.7}$$

As an example, we here summarize the main steps of the popular wavelet method in Thomas and Foken (2005), which is used for the comparison with the proposed method in Chapter 5 of this thesis. For each half hour interval of data, it detects structures following these main steps:

1. Normalize the given signal by subtracting the mean and dividing by the standard deviation.

2. Filter out low frequency motions from the high frequency turbulence using the BIOR5.5 wavelet function as a low pass filter, which consists of two sets of wavelets generated by a mother wavelet and a dual wavelet. The BIOR5.5 wavelet function has been shown to outperform HARR regarding localization in frequency (Kumar and Foufoula-Georgiou, 1994).

3. Determine the event duration (characteristic time scale) using the Morlet wavelet function according to the peak scale of the global wavelet spectrum. The Morlet wavelet function is defined as:

$$\psi(t) = \pi^{-1/4} e^{-i\omega_0 t} e^{-t^2/2}. \tag{1.2.8}$$

   It exhibits only one distinct peak when used to calculate the global wavelet spectrum (Thomas and Foken, 2005).

4. Detect individual structures by defining the zero-crossings of the Mexican-Hat wavelet coefficients at the chosen time scale as the moments of event occurrence (Collineau and Brunet, 1993a; Thomas and Foken, 2005). The Maxican-Hat wavelet is defined as the negative normalized second derivative of a Gaussian function $e^{-t^2/2}$:

$$\psi(t) = \frac{2}{\sqrt{3}\pi^{1/4}}(1 - t^2)e^{-t^2/2}. \tag{1.2.9}$$

Increasing- or decreasing-type structures can then be identified by observing the signs of the wavelet coefficients at the zero-crossing points.

1.2.5 Other Methods

Besides the above-mentioned event detection methods, some studies assume events have some types of geometric shapes and search for them in a supervised manner. The ramp function is most frequently used (e.g., Antonia and Fulachier, 1989; Wilczak, 1984; Chen *et al.*, 1997; Shapland *et al.*, 2012a,b; Barthlott *et al.*, 2007) while other patterns such as sine functions for waves or step functions for microfronts are also assumed (e.g., Mahrt, 2010; Belušić and Mahrt, 2012). Phase relationship has also been used as an indicator of structures (e.g., Campanharo *et al.*, 2008; Chian *et al.*, 2008).

1.2.6 A Summary of Event Detection Methods

As shown above, each event detection method focuses on certain components in the data. A summary of them is as follows.

- The existing methods search for structures with specific characteristics such as sharp gradients, large amplitudes, certain geometrical shapes, or high phase correlation. Variable Interval Time-Averaging (VITA) technique (e.g., Blackwelder and Kaplan, 1976; Collineau and Brunet, 1993b; Segalini and Alfredsson, 2012) looks for abrupt changes by calculating the short-time variance over a fixed time interval; Windowed Averaged Gradient (WAG) technique (e.g., Antonia and Fulachier, 1989; Bisset *et al.*, 1990; Collineau and Brunet, 1993b) detects rapid changes using a measure of the average gradient over some window length; the popular wavelet-based technique searches for large amplitudes with certain geometrical shapes resembling

the expected structure given by the chosen wavelet function (e.g., Farge, 1992; Collineau and Brunet, 1993a; Gilliam *et al.*, 2000; Chen and Hu, 2003; Thomas and Foken, 2005; Barthlott *et al.*, 2007; Segalini and Alfredsson, 2012). Other methods focus on different preassumptions of events.

- The VITA, WAG and quadrant techniques depend on the choices of several parameters which limits their application for an objective detection (Bogard and Tiederman, 1986).

- The objective wavelet-based methods perform better than others (e.g., Collineau and Brunet, 1993b; Hudgins and Kaspersen, 1999; Segalini and Alfredsson, 2012), but they are not good at distinguishing between signal and noise of comparable amplitudes and even tend to yield structures when provided with a noise-only time series (Collineau and Brunet, 1993a; Kang *et al.*, 2014c).

In conclusion, the existing event detection methods, which emphasize only a very limited number of large-amplitude or sharp-jump patterns, leave many other structures existing in the stable boundary layer not detected. Their origins, characteristics, and effects are currently unknown, although they could be an important contributor to the overall mixing and production of turbulence in stable conditions. In such situations, it is desirable to establish an event detection method that does not assume predefined geometries or other underlying physical processes.

1.3 Pattern Searching in Time Series Analysis

While the objective of detecting flow structures is to derive detailed physical properties of turbulence, an organization of the detected events into groups, if achievable, is essential to ease the analysis and understanding of typical dynamical

and behavioral patterns in stable ABL. From the perspective of time series data mining, event detection and classification in turbulence corresponds to extraction of previously unknown, frequently occurring patterns in time series.

An initial idea to achieve that is to cluster all the subsequences extracted using a sliding window and get the prototypes of the typical shapes in the series. However, Keogh *et al.* (2003) demonstrates that the clustering of overlapping subsequences of time series yields meaningless results since it always returns sine wave cluster centers regardless of the dataset. That triggered a great interest in the topic of motif discovery, which is proposed to make clustering of time-series subsequences meaningful. Motif discovery finds every subsequence (named motif) that approximately appears recurrently in a longer time series. This idea was transferred from gene analysis in bioinformatics. Since the definition of motif discovery was first introduced in Lin *et al.* (2002), several algorithms have been proposed (e.g., Mueen *et al.*, 2009b; Chiu *et al.*, 2003; Mueen *et al.*, 2009a; Yankov *et al.*, 2007; Wilson *et al.*, 2008; Lam *et al.*, 2011; Castro and Azevedo, 2012). In Lin *et al.* (2002), two user-defined parameters — a range R and a motif length n are specified. Two subsequences whose Euclidean distance is less than R forms a match. The most significant motif, known as 1-motif, is defined as the subsequence which has the largest number of non-trivial matches. Its generalization to k-motif returns the top-k motifs.

1.3.1 Pattern Searching Using Approximate Algorithms

The exact motif discovery solution for a time series presents high computational complexity (e.g. Lin *et al.*, 2005; Mueen *et al.*, 2009b; Castro and Azevedo, 2010). For that reason, most researchers have focused on approximate solutions to discover motifs, that can reduce the complexity by a large constant factor. Approximate algorithms search for motifs from discrete approximations of time series

instead of the raw time series. Lin *et al.* (2003) proposed a symbolic representation of time series, called SAX (Symbolic Aggregate approXimation), which provides opportunities for borrowing the already established techniques for finding the overrepresented DNA sequences in bioinformatics. For example, Chiu *et al.* (2003) proposed a probabilistic motif finding method based on random projection algorithm, which was successfully used in research for pattern discovery in biosequences. SAX representation of each time series subsequence is used to construct the base structure for the projection algorithm. It then hashes the subsequences into buckets. Buckets with multiple entries represent potential motif candidates. In Ferreira *et al.* (2006), time series subsequences were first transformed into the SAX representation before finding motifs in time series of proteins. In Tanaka *et al.* (2005), the authors introduced an algorithm to find motifs in multivariate time series. They firstly transform the multivariate time series into one signal using Principal Component Analysis. Motifs are found from the univariate data based on the algorithm in Chiu *et al.* (2003). Lin *et al.* (2005) used the SAX technique to generate a set of symbol strings for the subsequences from the time series, which are filtered into a suffix tree. The width of the tree branch shows the frequency of potential motifs. Wilson *et al.* (2008) proposed MTA (Motif Tracking Algorithm) using a novel immune inspired approach to evolve a population of trackers that seek out and match the motifs present in a time series. The first step of MTA is converting a time series to a symbolic representation using SAX. Castro and Azevedo (2012) presented an approach to calculate the statistical significance of the time series motifs. To derive each motif's significance p-value, they used iSAX (enhanced from SAX by Shieh and Keogh (2008)) to discretize the time series motifs.

1.3.2 Exact Pattern Searching

The above literature mainly uses approximate algorithms to find motifs. More recently, under the nearest-neighbor motif definition, Mueen *et al.* (2009b) proposed the first tractable exact motif discovery algorithm MK, to find exact motifs directly in the raw time series data. The motif search is made more efficient by early abandoning the Euclidean distance computation as soon as the cumulative sum goes beyond the current best-so-far and using the heuristic information calculated by the linear ordering of the distance of an object with respect to a few random reference points. MK, which is able to process very large datasets by up to three orders of magnitude faster than an exhaustive brute-force algorithm, is a sound contribution discussed a lot in the studies of motif discovery search (e.g., Mueen and Keogh, 2010; Lam *et al.*, 2011; Castro and Azevedo, 2010; McGovern *et al.*, 2011), so it is used in Chapter 4 of this thesis for a comparison with the proposed method.

1.3.3 A Summary of Pattern Searching

The potential application of the reviewed pattern searching algorithms in turbulence time series is summarized as follows.

1. The algorithms based on symbolic approximation are not advantageous for turbulence data. Besides their non-exact nature, approximate algorithms would introduce more difficulty in separating random noise from events when background random motions have comparable amplitudes than events, which can frequently happen in turbulence data.

2. The exact motif discovery algorithm is automatic and efficient. But we show later in the thesis that it tends to include a greater portion of noise in motifs

given the existence of the large amount of noise and the variety of noise types in turbulence data (Kang *et al.*, 2014c).

1.4 Objectives

Given a turbulence time series and assuming it is composed of non-random structures embedded in random noise, the goal of this thesis is to propose a novel method to extract and classify the structures from the time series and analyze their dynamics. From the above discussions, firstly, the initial idea of clustering sliding time series subsequences to get recurring structures is not suitable because of its meaninglessness (Keogh *et al.*, 2003). Secondly, the existing event detection methods in atmospheric sciences require assumptions on the event geometry or amplitudes, which is not advantageous for finding a larger variety of types of events (e.g., Antonia *et al.*, 1979; Wilczak, 1984; Collineau and Brunet, 1993a,b). Thirdly, the motif discovery algorithms in the time series data mining community tend to yield non-accurate events because of the ubiquity of noise in turbulence data (Kang *et al.*, 2014c).

To address these limitations, the following detailed objectives are set in this thesis.

1. Develop a method to detect and classify events from very noisy turbulence time series. The event detection method does not require any assumptions of characteristics of events in terms of their geometric properties. It is expected to be able to deal with the high level of background noise found in atmospheric turbulence time series. Also, it should be applicable for different atmospheric conditions and not limited to the scale of the phenomena. The event classification method needs to group the detected events into clusters with similar physical characteristics. Firstly, for easier physical interpretation of event clusters, the classification is expected to capture

the global characteristics of event time series so that events in each cluster have similar geometric shapes. Secondly, since the detected events do not necessarily have the same length, clustering should not assume time series with the same length.

2. Demonstrate the applicability of the proposed method — test the method on a well-known atmospheric turbulence time series, and validate the method by comparing application results to previous studies of the underlying physical mechanisms behind the dataset.

3. Apply the method to study the generally unknown stable ABL — detect the existing events, find the main physical and structural dynamics of different types of events and explore how these events affect atmospheric stability and turbulence.

4. Develop statistical software for the method and make it easily available for further analysis of ABL data.

1.5 Thesis Outline

This thesis addresses the first three objectives outlined above in five individual publications, which are included between the thesis introduction and conclusion chapters; and the fourth objective is addressed by a developed R software package, the manual of which is located in the thesis appendix. The contents in each chapter and their connections are described as below.

Chapter 1 of the thesis is an introductory chapter highlighting the motivations, difficulties of turbulence data analysis, discussion of literature and the objectives, which are briefly introduced above.

Chapters 2 to 4 address the first objective — method development. In Chapter 2 (Kang, 2012), we study unsupervised change detection in time series, by which we can see how detected patterns transit from one type to another. We propose a Growing Feature Quantization (GFQ) approach using a set of features to characterize a time series subsequence and then introduce a user defined parameter to control the growth rate of the cluster formation. Changes are defined as the transition of subsequences from one cluster to another. This method reveals transition migration of subsequences between existing clusters.

Chapter 3 (Kang *et al.*, 2014d) continues addressing the first objective. Despite the usefulness of GFQ in the synthetic and benchmarking data, it turns out to be not as applicable to turbulence data, due to the complexity of turbulence time series. This leads to two questions: (1) what are events? (2) what are their major characteristics? We start from the typical type of events that are frequently studied in literature — coherent structures. Note that other structures such as waves, microfronts, gravity currents, and other unknown phenomena can also be events. Detection of coherent structures assumes certain definitions of coherent structure, among which the most common one was given by Hussain (1983): "A coherent structure is a connected turbulent fluid mass with instantaneously phase-correlated vorticity over its spatial extent." Since then, the phase correlation is generally assumed to be a typical characteristic of structures, which could potentially be used for event detection. The main objective of this chapter is to study the relationship between coherent structures and their phase correlations. We first apply a standard wavelet-based technique to detect coherent structures from a real-world dataset (Thomas and Foken, 2005; Barthlott *et al.*, 2007), and examine their phase correlations using two different measures — the coherence index (Sahraoui, 2008; Chian *et al.*, 2008; Koga *et al.*, 2008; Hada *et al.*, 2003), and the nonlinearity measure based on nonlinear prediction error (nm_{npe}) (e.g., Sugihara and May, 1990; Schreiber and Schmitz, 1997; Rhodes and Morari, 1998;

Choudhury *et al.*, 2008b,a; Schreiber and Schmitz, 2000). It turns out that coherent structures do not necessarily have strong phase correlation. Therefore, using phase correlation as an indicator of coherent structures is not reliable.

Chapter 4 (Kang *et al.*, 2013) results from the uncertainty in the definition of structures and provides a big step to approach the first objective. In this chapter, considering the difficulty of directly quantifying characteristics of events, we shift our focus from finding events to firstly finding the noise in the time series since there is a large amount of noise in between the events. Since the focus of this study is on atmospheric time series, red noise is assumed for the background noise (Storch and Zwiers, 1999; Ghil *et al.*, 2002), although other types of noise are considered as well. Once noise is recognised and excluded, we define the remaining parts of the time series as potential events. A two-step method is proposed to realize the objective. As mentioned above, the first step is to extract events, by performing a noise test on each subsequence extracted from the series using a sliding window. All the subsequences recognized as pure noise are removed from further analysis, and the events are extracted from the remaining non-noise subsequences. The second step is to cluster the extracted events into similar patterns. This step is based on a set of features that carry the information about global characteristics of an event (Wang *et al.*, 2006). The proposed method, by firstly removing the noise subsequences and considering global characteristics rather than raw data, avoids the “meaningless subsequence clustering" limitation demonstrated in Keogh *et al.* (2003).

Chapter 5 (Kang *et al.*, 2014c) is to address the second objective. To make the method ready for further research, we verify it against a frequently studied real world ABL dataset — CASES-99 (Poulos *et al.*, 2002). One day of the 1-s averages of the thermocouple temperature data from CASES-99 is used for extraction, clustering and interpretation of events. Physical characteristics of events in each cluster are analyzed. Since a number of published studies have

examined the underlying physical mechanisms of several isolated events on that day (Thomas *et al.*, 2006; Wilczak, 1984; Williams and Hacker, 1992; Sun *et al.*, 2012; Blumen *et al.*, 2001; Poulos *et al.*, 2002), we test the performance of the method by comparing the results and insights with the previous studies.

Chapter 6 (Kang *et al.*, 2014a) is to address the third objective. The method is applied, with some improvements, on a more complicated turbulence dataset — FLOSSII (Mahrt, 2011b). This dataset is frequently characterized by the stable ABL, which has unknown structures. No published results can be found regarding extraction of previously unknown events. Apart from that, only a few studies have carried out a detailed and comprehensive analysis on the global physical characteristics of events in stable ABL. In this context, the purpose of this chapter is to find the events, classify them, characterize the dynamical and structural properties of different kinds of events, and especially discuss the contribution of events to atmospheric stability and turbulence.

Chapter 7 presents some concluding remarks with an overview of the results, the contributions of the thesis and directions for future work.

Appendix A (Kang *et al.*, 2014b), which comes last in this thesis, is to address the fourth objective. The event detection method was developed in R language (R Core Team, 2013). We have contributed the developed R package **TED** (Turbulence Event Detection and classification) to the Comprehensive R Archive Network (CRAN), a network of WWW sites holding the R distributions and contributed code, to share our work with the community. The manual of the package is appended.

Since this is a “Thesis by Publication" which consists of a new introduction and conclusion with published papers in between, unfortunately, it has inevitably created some amount of repetition among chapters. For the sake of thesis unity,

all the references of publications are located in a single Bibliography after chapter 7 and acknowledgements in the publications are covered by the thesis Acknowledgement.

PART B: Suggested Declaration for Thesis Chapter

Monash University

Declaration for Thesis Chapter 2

Declaration by candidate

In the case of Chapter 2, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
- Developed, established and verified the method - Wrote programming codes and the article	100

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature	[redacted]	**Date 21/05/2014**
Main Supervisor's Signature	[redacted]	**Date** 21/5/14

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Chapter 2

Real-time Change Detection in Time Series Based on Growing Feature Quantization

Chapter 2 is based on the article Kang Y. 2012. Real-time change detection in time series based on growing feature quantization. In: **Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN).** ***IEEE, pp. 1–6, doi: 10.1109/IJCNN.2012.6252381.***

Abstract. *An unsupervised time series change detection method based on an extension of Vector Quantization (VQ) clustering is proposed. The method clusters the subsequences extracted with a sliding window in feature space. Changes can be defined as transition of subsequences from one cluster to another. The method can be used to achieve real time detection of the change points in a time series. Using data on road casualties in Great Britain, historical data on Nile river discharges, MODerate-resolution Imaging Spectroradiometer (MODIS) Normalized Difference Vegetation Index data and x simulated data. It is shown that the detected changes coincide with identifiable political, historical, environmental or simulated events that might have caused these changes. Further, the online method has the potential for revealing the insights into the nature of the changes and the transition behaviours of the system.*

Keywords. *Change Detection; Feature Space; Vector Quantization; Time Series.*
References are considered at the end of the thesis.

Chapter 2

Real-time Change Detection in Time Series Based on Growing Feature Quantization

2.1 Introduction

Time Series data are generated, maintained, and processed within a broad of application domains in different fields. Mining such time series data becomes vital as the applications demand for understanding of the underlying processes or phenomena that generate the data. A specific interesting mining task is to detect changes in a given time series. Early identification of changes in a time series is one of the most promising topics in statistics (Verbesselt *et al.*, 2010a; Zeileis *et al.*, 2003; Bai and Perron, 2003; Verbesselt *et al.*, 2010b) and data mining (Sharifzadeh *et al.*, 2005; Kifer *et al.*, 2004; Firoiu and Cohen, 2002; Keogh *et al.*, 2001b; Tsai and Shieh, 2009) due to the numerous applications where early warning systems are needed. Also, known as change detection or sometimes event detection, this problem covers a broad range of areas of application including land cover

change detection (Verbesselt *et al.*, 2010a; Salmon *et al.*, 2011; Lunetta *et al.*, 2006), early warning of pandemic outbreaks (Culotta, 2010), signal segmentation in data streams (Keogh *et al.*, 2001b; Firoiu and Cohen, 2002), fault detection in engineering systems (Fujimaki *et al.*, 2005), telecommunication network (Burge *et al.*, 1997), economics (Jouini and Boutahar, 2005) and business (Tsai and Shieh, 2009).

The aim of this paper is to propose a new method for real-time change detection, that generates insights into the transition behaviours of the system. Vector Quantization (VQ) is a popular and widely applied clustering algorithm (Gray, 1984), which moves clustering centres denoted as code-book vectors towards accumulation points in the data set. The algorithm will be described in section 2.3.2. We propose a growing VQ approach using a set of features to characterise a time series subsequence and then introduced a user defined parameter to control the growth rate of the cluster formation. Then changes are defined as the transition of subsequences from one cluster to another. This method reveals transition migration of subsequences between existing clusters and helps find new states of the system.

The paper is organised as follows. Section 2.2 presents a brief review of relevant research. Section 2.3.1 represents the feature extraction of sliding windows. Section 2.3.2 presents vector quantization clustering and change detection based on a new algorithm which we call growing feature quantization (GFQ). Section 2.3.3 describes how to recognise changes and transitions between clusters. Section 2.4 presents the results for three well-known datasets and two simulated time series. Section 2.5 presents the conclusions.

2.2 Related Work

A typical statistical formulation of change-point detection is to consider probability distributions from which data in the past and present intervals are generated, and regard the current time point as a change point if two distributions are significantly different (Kawahara and Sugiyama, 2009). Various other approaches have been investigated, such as the CUSUM (CUmulative SUM) (Zeileis *et al.*, 2003; Verbesselt *et al.*, 2010a), direct density-ratio estimation (Kawahara and Sugiyama, 2009) and unsupervised time series subsequence clustering (Salmon *et al.*, 2011). CUSUM detects changes by investigating the sum of linear regression errors. When the errors exceed a threshold, we consider that the time series no longer fits the regression model and a change occurred. Direct density-ratio estimation is a non-parametric approach to estimate the ratio of probability densities. Whether there is a change point is decided by monitoring the logarithm of the likelihood ratio. The unsupervised time series subsequence clustering clusters the subsequences and defines the transition of the subsequence from one cluster to another as a change. However, those approaches only give indication if change has occurred rather than providing insights into the nature of the change and the transition behaviour of the system. In Salmon *et al.* (2011), the unsupervised clustering method to detect land cover change has the potential for revealing patterns in the system, but it could not be used to deal with recently acquired data.

Considering that a rapid response or early warning is crucial in many cases, this paper proposes a method for real time detection of the change points in time series. The proposed method is based on time series subsequence clustering. There are two main categories in time series clustering (Keogh *et al.*, 2003). “Whole clustering" is similar to that of conventional clustering of discrete objects. The entire time series is taken as a discrete object. In contrast, “subsequence clustering"

is performed on individual subsequences extracted with a sliding window. A subsequence $x_p(t)$ for a time series $x(t)$ with length m is

$$x_p(t) = (x(t_p), \cdots, x(t_{p+w-1})) \tag{2.2.1}$$

for $1 \leq p \leq m - w + 1$, where w is the length of the subsequence. The sequential subsequences in (1) are extracted using a sliding window with a length of w and position p, which is incremented with a natural number $\mathbb{N}$. Widespread use of subsequence clustering has been made in different areas. However, the sliding window causes the clustering procedure to create meaningless results as it forms sine wave cluster centres regardless of the data set, which makes the clusters extracted by any clustering algorithm essentially random (Keogh *et al.*, 2003). To address this problem, several solutions have been used. Keogh *et al.* (2003) demonstrated a meaningful motif-based-clustering method. Chen (2005) and Goldin *et al.* (2006) used alternative distance measures to make sequential time series clustering meaningful. In Wang *et al.* (2006), global measures describing time series were proposed to capture the underlying characteristics: trend, seasonality, periodity, serial correlation, skewness, kurtosis, chaos, nonlinearity and self-similarity, and the clustering was performed on the subsequences defined by a feature vector of these measures. Salmon *et al.* (2011) demonstrated three different unsupervised clustering approaches that operate on short term Fourier transform coefficients computed over subsequences that are extracted with a temporal sliding window and created meaningful sequential time series. Here we borrow the idea of Salmon *et al.* (2011) and Wang *et al.* (2006) and use a set of subsequence features to map the original subsequences into feature space before clustering subsequences meaningfully. However, changes are detected here in an on-line manner while Salmon *et al.* (2011) operates clustering off-line.

2.3 Proposed Methodology

2.3.1 Feature Extraction

It is claimed in Keogh *et al.* (2003) that non-overlapping sliding windows, with their positions incremented by exactly the periodic length, would produce valid clusters when applied to a periodic time series. However, using the magnitude of the first few Fast Fourier Transform (FFT) components of $x_p(t)$ to characterise the subsequences makes the sliding window position p not have to be shifted by a fixed amount, but can be incremented by any natural number $\mathbb{N}$ (Salmon *et al.*, 2011). For each subsequence $x_p(t)$, the features $x_p(f)$ are computed as

$$X_p(f) = |\mathcal{F}(x_p(t))| \tag{2.3.1}$$

where $\mathcal{F}(\cdot)$ represents the Fourier transform. The window length w depends on the type of time series. For seasonal time series, w is always fixed at the number of samples corresponding to the length of the cycle.

Moreover, additional features beyond FFT components like chaotic properties, serial correlation and so on (Wang *et al.*, 2006) could be calculated to characterize the time series subsequences.

2.3.2 Unsupervised Change Detection: Growing Feature Quantization

VQ clustering is a classical quantization technique to divide a large set of points (vectors) into groups. Each group is represented by its centroid point. Its goal is to discover structure in the data by finding how the data is clustered. In VQ, there is a codebook which is defined by a set of M prototype vectors. M is chosen by the

user and the initial prototype vectors are chosen arbitrarily. An input belongs to cluster i if i is the index of the closest prototype. From the mathematical point of view, vector quantization is basically a simplified version of k-means (Lughofer, 2008). The simple idea is in Algorithm 1.

Algorithm 1 VQ

1: Choose the number of clusters M
2: Initialize the prototypes $w_1, \cdots, w_m$
3: Randomly pick an input x
4: Find the winning cluster w^* by finding the prototype vector satisfying

$$|w^* - x| \leq |w_i - x|, i = 1, \cdots, M \tag{2.3.2}$$

5: Update the winning prototype weights according to

$$\bar{w^*}_{new} = \bar{w^*}_{old} + \eta * (x - \bar{w^*}_{old}) \tag{2.3.3}$$

where η is the adaptation value

Algorithm 1 can not be applied for data sets with an unknown number of clusters. Various clustering approaches have been presented in an incremental manner such as sequential k-means (Duda *et al.*, 2001), dynamic Self Organised Maps (SOM) (Alahakoon *et al.*, 2000) and Growing Neural Gas (Jirayusakul and Auwatanamongkol, 2007; Sledge and Keller, 2008). The GFQ clustering is proposed in this paper. The goal is to cluster the subsequence features incrementally, by which new clusters can be recognized in time and the number of clusters do not have to be known in advance. In some systems like infectious diseases the earliest possible warning of a change is required, while in other systems an early warning of changes costs a lot of energy, money and sometimes panic. To enable the sensitivity of the system to be controlled, a user defined single parameter R is used. For each incoming feature vector x, if this condition is fulfilled:

$$\text{distance}(x, w^*) \geq R \tag{2.3.4}$$

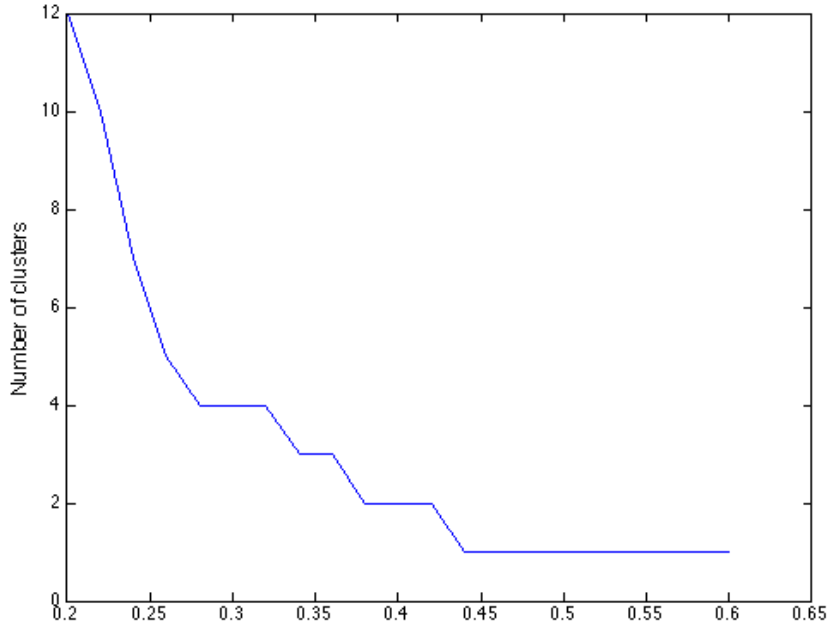

Figure 2.1: *How the choice of R influences the number of clusters*

where w^* is the winning prototype, we create a new cluster, which x belongs to. Otherwise, x belongs to the winning cluster. The number of clusters will become smaller with the parameter R growing (Fig. 2.1). This parameter, determined by trial and error, should be around $\sqrt{d}/3$, where d is the dimension of the feature space. Of course, this parameter can be flexibly tuned according to the real world context to reduce false alarms or increase early warnings. The whole process is summarized in Algorithm 2.

A one-pass incremental and evolving variant of VQ were demonstrated in Lughofer (2008) by incorporating a vigilance parameter, exploiting the idea in the adaptive resonance theory (ART) (Carpenter and Grossberg, 2010). However the prototype vectors are not rescaled when the incoming input is outside the estimated range, which actually places the new input on a different scale to the prototype vectors.

Algorithm 2 GFQ

1: Initialize a threshold R, which gives a radius around a cluster prototype, in which feature vectors must lie to belong to the cluster
2: Initialize an adaptation value η, which depends on the number of inputs in the cluster
3: Collect a few data samples to obtain the estimated maximum and minimum for each feature component, and hence the estimated ranges of each feature
4: Initialize $C = 1$, where C is the current number of clusters; initialize a cluster prototype w_1, which is the first normalized input
5: Read the next incoming subsequence and calculate its feature vector x as the new input
6: **if** x is outside the estimated range **then**
7: Update the ranges of each feature
8: Rescale the current cluster prototypes using the updated ranges of each feature
9: **else**
10: Use the current estimated ranges of each feature
11: Normalize the input to $[0,1]^d$ according to the ranges, where d is the dimension of the feature space. Name the normalized input as $\hat{x}$
12: Find the winning cluster w^* by finding the prototype vector satisfying

$$|w^* - \hat{x}| \leq |w_i - \hat{x}|, i = 1, \cdots, C \tag{2.3.5}$$

13: **if** distance$(\hat{x}, w^*) < R$ **then**
14: Make $\hat{x}$ a member of w^*
15: Let $C = C$
16: Update the winning cluster center:

$$\bar{w^*}_{new} = \bar{w^*}_{old} + \eta * (\hat{x} - \bar{w^*}_{old}) \tag{2.3.6}$$

17: Update the adaptation rule:

$$\eta = 1/(\text{number of inputs in the cluster}) \tag{2.3.7}$$

18: **else**
19: Create a new cluster; make $\hat{x}$ a member (and the center) of the new cluster
20: Let $C = C + 1$

2.3.3 Recognising changes and transition behaviours

Changes in time series are defined as the transition of the subsequence from one cluster to another within the feature space which characterise the time series subsequences. Thus the transition behaviours can be identified according to the memberships of the subsequences. It reveals whether the transition is between existing clusters or it changes to a new state. In this way, the states of the system can be both qualitatively and quantitatively described.

2.4 Results

2.4.1 The seatbelt data

The seatbelt data is a monthly time series (from Jan 1969 to Dec 1984) of the number of car drivers in Great Britain killed or seriously injured in traffic accidents. There are two breakpoints in this time series, which are Oct 1973—associated with petrol rationing and the introduction of lower speed limits during the first oil crisis—and Jan 1983—associated with the seat belt law introduced in the UK on 1983-01-31 (Harvey and Durbin, 1986). The sliding window length is fixed at $w = 12$ samples to correspond to the length of the annual cycle. We use magnitudes of the first four FFT components to characterize the sliding windows. Global features like chaotic properties are not used here because the short length of the subsequences and the short sliding step will not make those features of the subsequences well distinguished. In Fig. 2.2, the circles represent the ending points of the corresponding subsequences. Different colors means the subsequences are grouped into different clusters. From Fig. 2.2, using GFQ clustering, the transitions from one cluster to another can be seen both in end of 1973 and beginning of 1983.

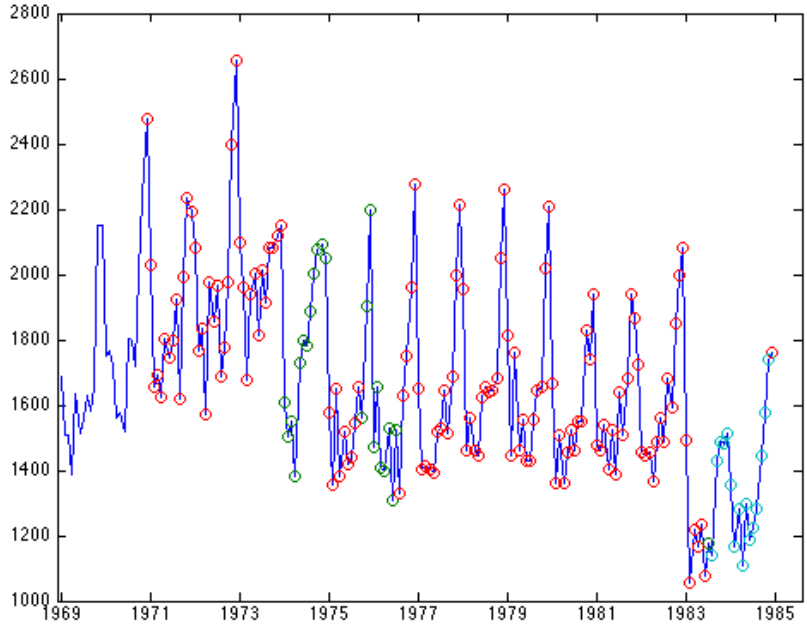

Figure 2.2: *SeatBelt time series*

To reveal the transition process in this system, denote the three clusters in this system as states $S1, S2, S3$. At the end of 1973, the system changed from $S1$ to $S2$, after which the system went back to state $S1$ from the beginning of 1976 till 1983. After that, there came a new state $S3$ in the beginning of 1983 because of the introduction of the seat belt law.

To make comparisons, Fig. 2.3 gives the clustering results using growing vector quantization based on raw data (but not features). The results are in line with the "meaningless " interpretation reported in Keogh *et al.* (2003). On the other hand, Fig. 2.2 indicates the meaningful subsequence clustering based on features. Fig. 2.4 gives the clustering results using k-means based on features. The two main changes identified using GFQ are roughly in accordance with the changes detected using k-means. This indicates that GFQ can obtain similar quality results as k-means but GFQ clustering is a real-time method, in which the number of clusters don't have to be known in advance.

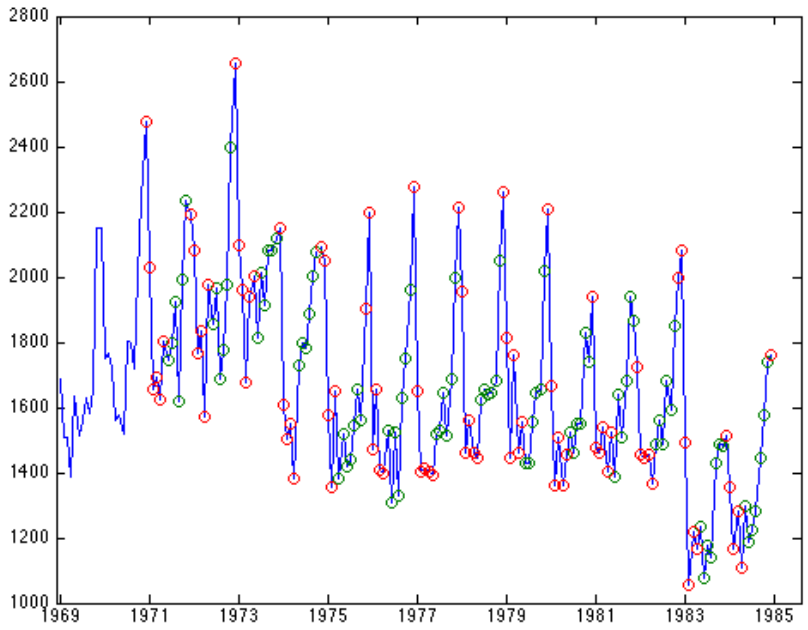

Figure 2.3: *SeatBelt time series subsequence clustering using growing vector quantization based on raw data*

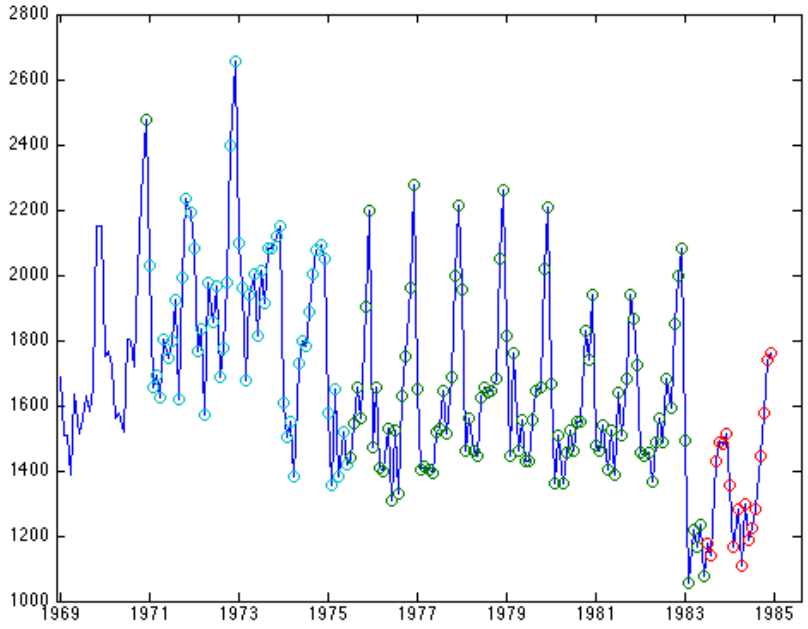

Figure 2.4: *SeatBelt time series subsequence clustering using k-means based on features*

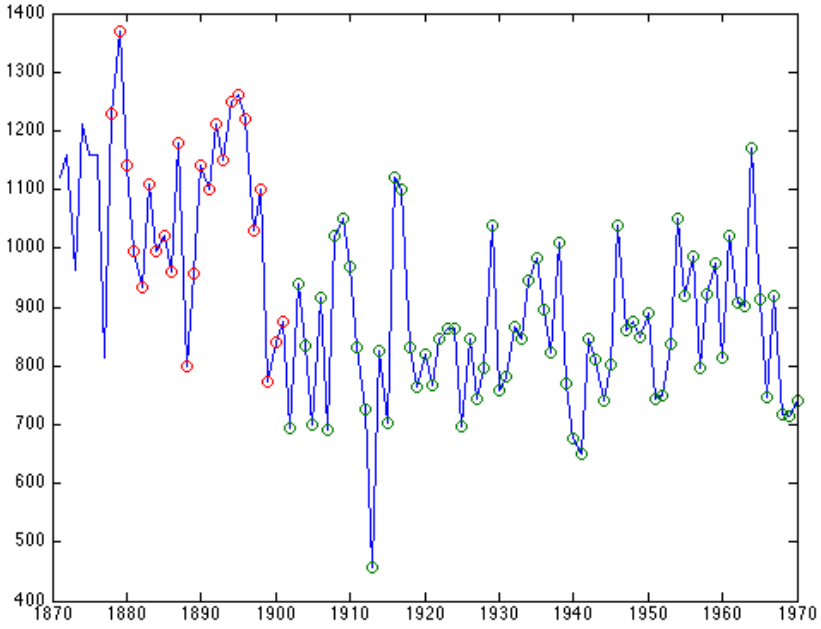

Figure 2.5: *Nile time series with changes using GFQ*

2.4.2 The Nile data

The Nile data is a time series of the annual flow of the river Nile at Aswan from 1871 to 1970 (Zeileis *et al.*, 2003; Cobb, 1978). It measures annual discharge at Aswan in $10^8 m^3$. From Fig. 2.5, we can see that there is a change around 1900. The obvious reason is the Aswan dam that was built in 1898. Fig. 2.6 shows the distance from the data points to the winning prototype. From 1871, the data points are moving further from the first prototype until a new cluster is created around 1900 when the distance of the incoming data points to the original prototype exceeds the pre-defined threshold R.

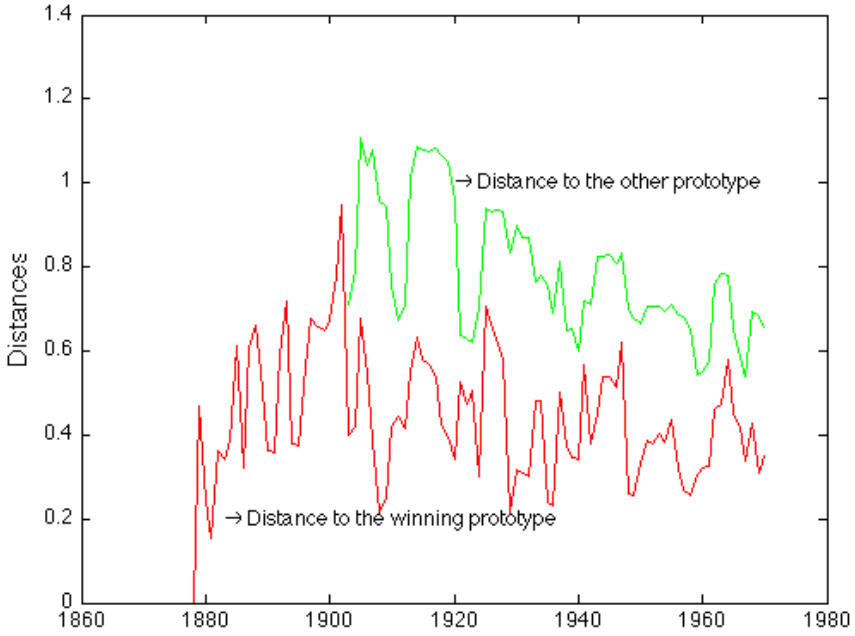

Figure 2.6: *Distances to prototypes*

2.4.3 The MODIS NDVI data

The MODIS NDVI data is an NDVI time series of a *pinus radiata* plantation (Verbesselt *et al.*, 2010a). The transition of clusters can be seen in Fig. 2.7 around the year 2004, which is the known harvest year.

2.4.4 Simulated time series

Simulated time series are generated by summing individually simulated seasonal, noise and trend components (Verbesselt *et al.*, 2010a). The seasonal component is created using an asymmetric Gaussian function for each season, which has been shown to perform well when used to extract seasonality (Jonsson and Eklundh, 2002). The noise component is generated using a random number generator that follows a normal distribution. The trend component was generated by selecting a constant. Two drops and linear recovery phases in trend component are simulated

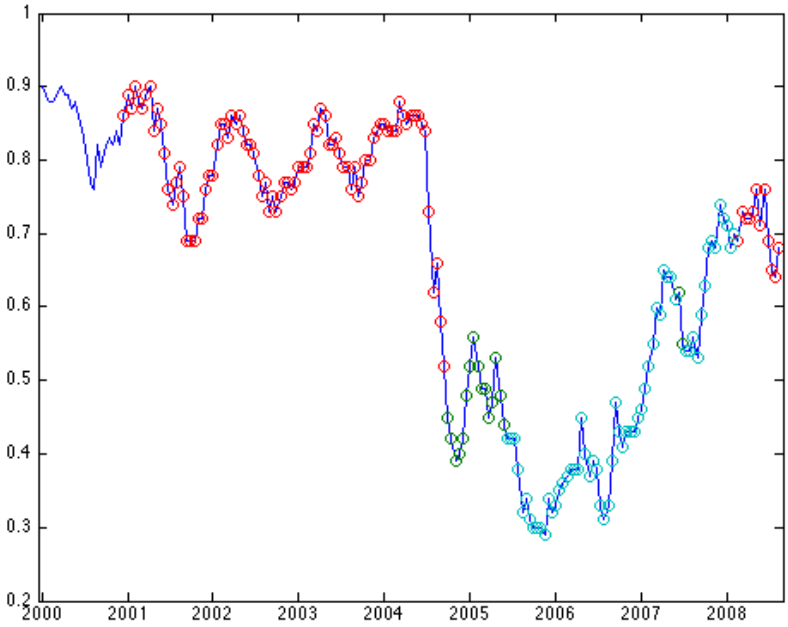

Figure 2.7: *MODIS time series*

in Fig. 2.8. An additional change in seasonal component is simulated in Fig. 2.9. From the transition of the subsequence membership, the simulated changes are detected easily.

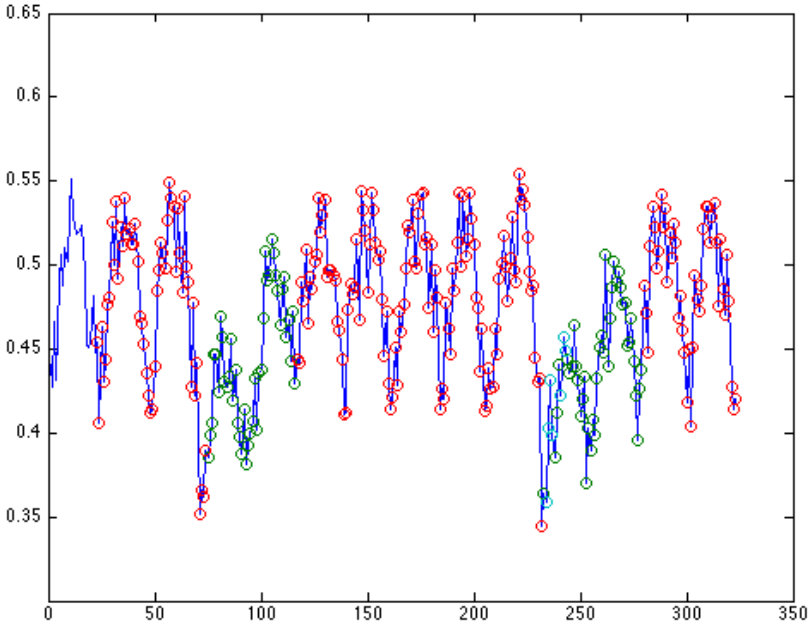

Figure 2.8: *Simulated time series*

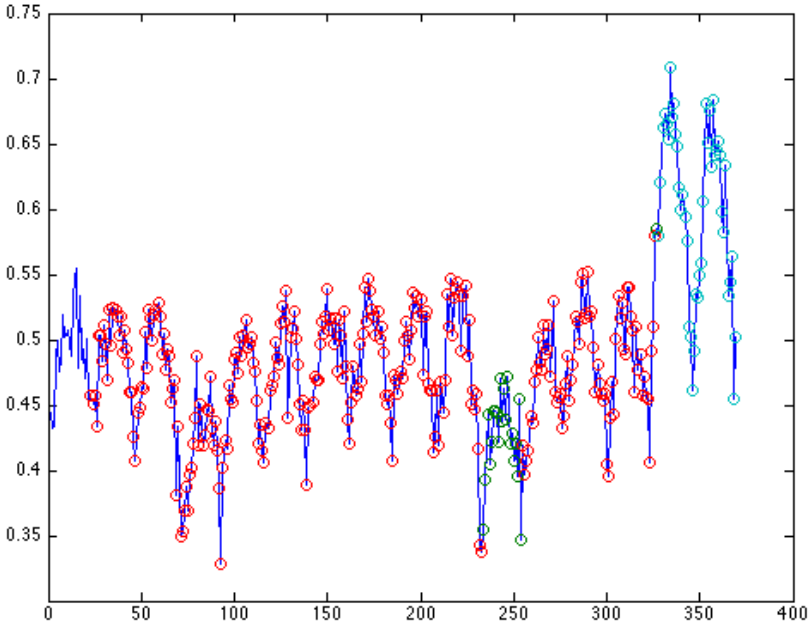

Figure 2.9: *Simulated time series*

2.5 Conclusion

In this paper, a method for real-time time series change detection is proposed. It is based on an extension of VQ—known as growing feature quantization (GFQ) clustering, which provides insights into the transitions of the time series subsequence states. In order to avoid the meaninglessness limitation pointed out in Keogh *et al.* (2003), the method uses features instead of raw data to characterise the time series subsequences. According to the experiments on three frequently used time series as well as two simulated data, the proposed approach performs as well as k-means, but it can be used to detect changes in a real-time manner and the number of clusters don't have to be known in advance. In addition, the method can reveal the transitions among the system, provide insights into the nature of the pattern changes and find new states coming in the current system.

Further research is necessary to study the choice of subsequence features. Extension of features to a more comprehensive feature set will be studied. Future work also involves the choice of the threshold R. This parameter can be flexibly tuned according to the real world context to reduce false alarms or increase early warnings. That means it depends on trade-off between the benefits of early warning and the misclassification costs in the system. It is necessary to find an optimal threshold R^* to detect changes reasonably for a specified system. Besides, larger datasets in more fields of application such as sleep staging (Stanus, 1986; Acharya *et al.*, 2010; Lee *et al.*, 2009; Güneş *et al.*, 2010) will be tested using GFQ change detection approach.

PART B: Suggested Declaration for Thesis Chapter

Monash University

Declaration for Thesis Chapter 3

Declaration by candidate

In the case of Chapter 3, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
- Developed, established and verified the method - Wrote programming codes and the article	90

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Danijel Belusic	Provided helpful guidance and editorial work	
Kate Smith-Miles	Provided helpful guidance and proofreading	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature	[redacted]	**Date** 21/05/2014
Main Supervisor's Signature	[redacted]	**Date** 21/5/14

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Chapter 3

A Note on the Relationship between Turbulent Coherent Structures and Phase Correlation

Chapter 3 is based on the article Kang Y, Belušić D, Smith-Miles K. 2014d. A note on the relationship between turbulent coherent structures and phase correlation.* Chaos: An Interdisciplinary Journal of Nonlinear Science *24(2): 023114, doi: http://dx.doi.org/10.1063/1.4875260.

Abstract. *Various definitions of coherent structures exist in turbulence research, but a common assumption is that coherent structures have correlated spectral phases. As a result, randomization of phases is believed, generally, to remove coherent structures from the measured data. Here we reexamine these assumptions using atmospheric turbulence measurements. Small-scale coherent structures are detected in the usual way using the wavelet transform. A considerable percentage of the detected structures are not phase correlated, although some of them are clearly organized in space and time. At larger scales, structures have even higher degree of spatio-temporal coherence but are also associated with weak phase correlation. A series of specific examples are shown to demonstrate this. These results warn about the vague terminology and assumptions around coherent structures, particularly for complex real-world turbulence.*

References are considered at the end of the thesis.

Chapter 3

A Note on the Relationship between Turbulent Coherent Structures and Phase Correlation

3.1 Introduction

The complexity of turbulent flows is sometimes addressed by studying their deterministic coherent structures, which have significant influence on the flow and the random background motions (Ouellette, 2012). As a result, identifying coherent structures from turbulent flows, which is akin to distinguishing order from disorder, has been a popular topic in turbulence research. This has resulted in a number of methods for their detection, all of which assume certain definitions of a coherent structure. Perhaps the most common assumption is that coherent structures are phase correlated. The relationship between coherent structures and phase correlation dates back to the early works on coherent structures, when Hussain (1981, 1983, 1986) defines a coherent structure as “a connected turbulent fluid mass with instantaneously phase-correlated vorticity over its spatial extent".

Even though Hussain (1986) writes "In principle, concepts like coherent structures are best left implicit", he still gives the definition to facilitate measurements and analysis of coherent structures. Measurements of real world turbulence, such as in the atmospheric boundary layer, are predominantly in-situ and result in single-point time series. The spatial information is missing in those cases, and raw turbulence variables in time domain are assumed to be phase correlated instead. The lack of spatial information also limits the use of recent techniques focused on the Lagrangian characteristics of coherent structures, which seem to be more suitable for detecting coherent structures than the Eulerian techniques (Peacock and Dabiri, 2010; Tang *et al.*, 2010; Tang and Peacock, 2010; Miranda *et al.*, 2013).

It appears that since the definition was given, it has become commonly assumed that the phase correlation is a necessary characteristic of a coherent structure. A typical example is the phase randomization technique, which has been applied in many dynamical systems for examining their non-linearity (e.g. Jeong *et al.*, 2002; Lan *et al.*, 2005; Guarin-Lopez *et al.*, 2010; Kugiumtzis, 2002; Waser, 2010; Pereda *et al.*, 1998), testing for chaos (Gomes *et al.*, 2000; Lan *et al.*, 2005; Campanharo *et al.*, 2008; Lin, 2005; Jeong *et al.*, 2002), or removing coherent structures (e.g. Belušić and Mahrt, 2012; Arzner *et al.*, 2006; Campanharo *et al.*, 2008; Chian *et al.*, 2008; Koga *et al.*, 2008; Sahraoui, 2008). It has been shown to work well for a variety of applications, ranging from simple dynamical systems, such as the Lorenz attractor (Provenzale *et al.*, 1992), to more complex astrophysics and atmospheric fluid flows (e.g. Arzner *et al.*, 2006; Campanharo *et al.*, 2008). A literature review of applications of phase randomization indicates that it is regarded as a tool that can be universally applied, with no reports of its limitations. However, Belušić and Mahrt (2012) find that in complex atmospheric flows, the phase randomization does not remove all structures. While this is not surprising, since the atmosphere contains myriads of structures with different characteristics and

generating mechanisms, it calls into question the assumed universal relationship between coherent structures and phase correlation.

The relationship between coherent structures and phase correlation is examined in two ways. We first apply a standard wavelet-based technique to detect coherent structures from a real world dataset. Their phase correlation is examined using two measures — the coherence index (Sahraoui, 2008; Chian *et al.*, 2008; Koga *et al.*, 2008; Hada *et al.*, 2003), and the nonlinearity measure based on nonlinear prediction error (nm_{npe}) (e.g. Sugihara and May, 1990; Schreiber and Schmitz, 1997; Rhodes and Morari, 1998; Choudhury *et al.*, 2008b,a; Schreiber and Schmitz, 2000). Both measures are obtained from a comparison between statistics of the original time series and its phase-randomized surrogates, which is a general approach for detecting nonlinearity (Theiler *et al.*, 1992; Schreiber and Schmitz, 2000). The coherence index uses the first order structure function as the statistic, while nm_{npe} uses the nonlinear prediction error. The two different measures are used in order to cross-validate the results. The details of the methods are discussed in Section 3.2.3 and Section 3.2.4.

The second way is to examine specific examples of coherent structures that are coherent in space and time, but are not phase correlated. The spatio-temporal coherence of structures is examined from the time-height cross-sections of turbulent variables and their relationships. Results in Section 3.3 show that some physical coherent structures in turbulent flows indeed have weak phase correlation. This may be important for analyses where the effects of coherent structures are examined by removing them from the original data.

3.2 Data and Methods

3.2.1 Dataset Description

The data are from the CASES-99 field experiment (Poulos *et al.*, 2002). The measurements were performed during October 1999 on a 60-m tower located at a rural grass site near Leon, Kansas, USA. The tower was equipped with thermocouples and sonic anemometers at several levels. The thermocouples sampled data at 5 Hz and were located at 34 vertical levels (0.23 m, 0.63 m, 2.3 m, and every 1.8 m above 2.3 m) (Sun *et al.*, 2012), while the 20-Hz sonic data were measured at seven vertical levels (1.5, 5, 10, 20, 30, 40, 50, and 55 m). A detailed description of the CASES-99 experiment can be found in Poulos *et al.* (2002). Here we use the thermocouple and sonic anemometer data downsampled to 2 Hz.

3.2.2 Wavelet Detection of Coherent Structures

Detecting coherent structures from time series using wavelets is the most frequent approach in the atmospheric turbulence research. It has been shown to outperform other classical approaches, such as the Variable Interval Time-Averaging (VITA) (Blackwelder and Kaplan, 1976) and the Windowed Averaged Gradient (WAG) (Bisset *et al.*, 1990). We use the method proposed by Thomas and Foken (2005). It is an automated and quasi-online coherent structure detection algorithm based on wavelets, which first high-pass filters the time series by using the bi-orthogonal wavelets BIOR5.5, then uses the Morlet wavelet to determine the characteristic scales of structures, and finally detects the individual structures using the Mexican hat wavelet. Following Barthlott *et al.* (2007), we additionally apply a threshold of 40% of the absolute maximum of the wavelet coefficients at the characteristic scale to reduce the number of false detections.

3.2.3 Estimation of Coherence Index

A general index of coherence $CI(q,\tau)$ (Sahraoui, 2008) is used to estimate the degree of phase correlation among Fourier modes of a time series. $CI(q,\tau)$ evaluates the standardized difference between the structure functions of the original time series $x_O(t)$ and its phase randomized surrogate $x_R(t)$. Phase randomization randomly scrambles the Fourier phases of the time series while keeping the spectral magnitudes unchanged. The surrogate time series $x_R(t)$ is a stationary Gaussian linear process obtained after taking the inverse Fourier transform of the phase-randomized spectrum. The completely coherent surrogate $x_C(t)$ is used for standardization of the difference. $x_C(t)$ is obtained by making the phases constant. Formally,

$$CI(q,\tau) = \left(\frac{|S_O(q,\tau) - S_R(q,\tau)|}{|S_O(q,\tau - S_R(q,\tau))| + |S_O(q,\tau) - S_C(q,\tau)|} \right)^{1/q}, \tag{3.2.1}$$

where $S_i(q,\tau)$ is the qth order structure function: $S_i(q,\tau) = \langle |x_i(t+\tau) - x_i(t)|^q \rangle$, $i \in \{\text{O,R,C}\}$ and τ is the time lag. It has been shown that different orders ($q = 1, 3, 4$) yield mainly similar coherence index (Sahraoui, 2008), so we use $q = 1$ to simplify the calculations. With this choice of q, our coherence index becomes comparable to the index used in Chian *et al.* (2008), Koga *et al.* (2008), Hada *et al.* (2003). The coherence index values range from 0 to 1, where 0 indicates the original time series with random phases, and 1 with completely correlated phases. The maximum τ considered here is 1/4 of the time series length, as suggested by Sahraoui (2008).

3.2.4 Nonlinearity Measure Based on Nonlinear Prediction Error (nm_{npe})

The nonlinearity measure nm_{npe} compares the predictability of a time series $x(t)$ with its phase randomized surrogates. Nonlinear time series with phase correlation are more predictable than their surrogates, and thus have smaller nonlinear prediction errors (Choudhury *et al.*, 2008b; Schreiber and Schmitz, 2000, 1997). Nonlinear prediction uses an embedding matrix X consisting of delay vectors ($\vec{x}_i$, $i = 1, 2, \cdots, n-m+1$) as the rows in m dimensions:

$$X = \begin{bmatrix} x(1) & x(2) & \cdots & x(m) \\ x(2) & x(3) & \cdots & x(m+1) \\ \cdots & \cdots & \cdots & \cdots \\ x(n-m+1) & x(n-m+2) & \cdots & x(n) \end{bmatrix}$$

For each row (delay vector) of X, its k nearest m-dimensional delay vectors are found using Euclidean distances. If the k nearest delay vectors for $\vec{x}_i$ are $\vec{x}_{j_p}$, $p = 1, 2, \cdots, k$, the nonlinear prediction error for the time series $x(t)$ is defined as

$$\mathcal{T}_X(m,k) = \sum_{i=1}^{n-m} \left(x(i+m) - \frac{1}{k} \sum_{p=1}^{k} x(j_p+m) \right)^2 . \tag{3.2.2}$$

The nonlinearity measure for the time series $x(t)$ is obtained from

$$nm_{npe} = \frac{\bar{\mathcal{T}}_R - \mathcal{T}_X}{3\sigma_R}, \tag{3.2.3}$$

where $\mathcal{T}_X$ is the nonlinear prediction error of the original data, while $\bar{\mathcal{T}}_R$ and σ_R are the mean and standard deviation of the nonlinear prediction errors of the surrogates. Following Choudhury *et al.* (2008b), we use $k = 8$ and generate 50

surrogates to get the statistical distribution of the measures. The maximum m considered here is 1/4 of the time series length, which is consistent with the maximum of τ in Section 3.2.3. For nonlinear time series, we have $nm_{npe} \geq 1$, while when $nm_{npe} < 1$, the time series is considered to be phase-uncorrelated.

3.3 Results

3.3.1 Wavelet-Detected Coherent Structures

Wavelet analysis is applied to each 30 min of the thermocouple temperature time series at the seventh level (9.5 m), from 1100 LST 5 October 1999 to 1100 LST 6 October 1999. Similar results can be obtained for other days. In total, 252 structures are found with different event duration, and the coherence index is calculated for each identified structure. Fig. 3.1(a) shows the distribution of the maximum coherence index over the considered lags τ (see Section 3.2.3) for the detected coherent structures. Even though these structures are termed "coherent", about 25% of them have the maximum coherence index smaller than 0.3 and 59% smaller than 0.5. These coherent structures are deemed to have weak phase correlation. Fig. 3.1(b) shows the distribution of the maximum nm_{npe} over the considered embedding dimension m (see Section 3.2.4). About 72% of the events are not phase-correlated according to this measure. Some of the phase-uncorrelated events are shown as examples in the next subsection.

3.3.2 Examples of Phase-Uncorrelated Coherent Structures

A number of examples of space and time coherent structures with weak phase correlation are presented. Fig. 3.2 shows two coherent structures detected from the CASES-99 temperature time series. The coherence index of the structures is

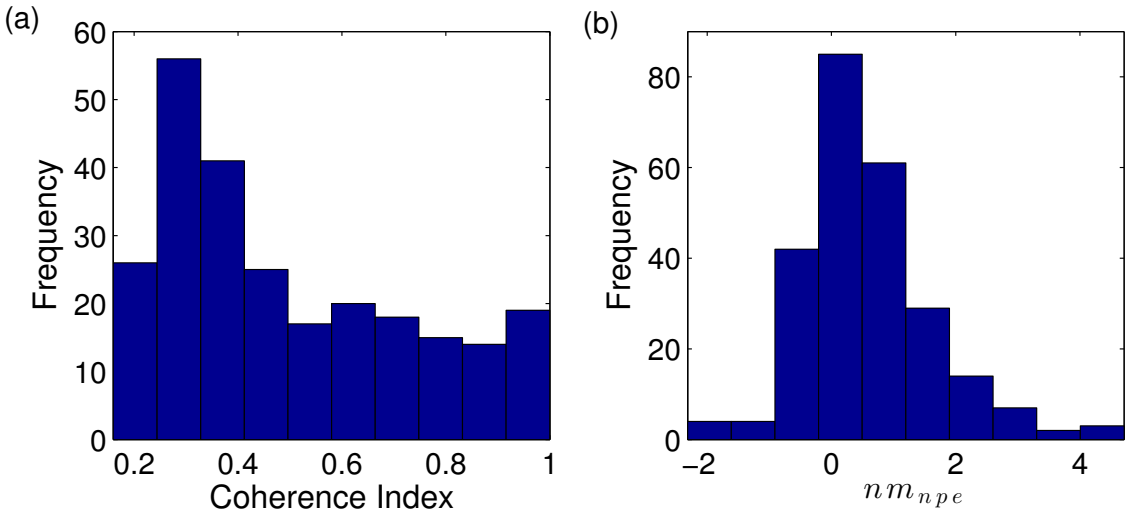

Figure 3.1: *Frequency distributions of (a) the maximum coherence index over the considered lags and (b) the maximum nonlinearity measure nm_{npe} for the coherent structures extracted from the CASES-99 data using wavelets.*

below 0.3 for all considered lags τ. The maximum nm_{npe} values are smaller than 1 for both events, which confirms that they are linear processes without phase correlation. The time-height vertical cross-sections of temperature and horizontal wind speed (Fig. 3.3) show that both structures are vertically coherent over the distance of at least 50 m. The temperature and horizontal wind speed are out of phase, indicating that the structures are organized motions in convectively unstable boundary layer turbulence. In such conditions, upward turbulent motion results in negative speed perturbations and positive temperature perturbations, and vice versa. This relationship is confirmed for the current cases by examining the relative phase angles between the vertical wind speed and temperature or horizontal wind speed (not shown).

Turbulence in the atmosphere occurs on a wide range of scales, and the energy and spatial scale of organized structures generally increase with time scale. We therefore expect to observe structures with higher degree of coherence over the measurement tower height at larger scales. Here we show two examples from the same dataset with about ten times larger time scales, on the order of 10 min. These structures are not extracted using the current wavelet method, because

it is designed to detect structures with time scales of the order of 1 min, which is typical in atmospheric science applications. Fig. 3.4 shows the two coherent structures over a 30-min period. The coherence index is small, indicating that these structures have low phase correlation. The latter is confirmed by the maximum nm_{npe} values less than 1. However, the time-height cross sections (Fig. 3.5) show that they are organized, vertically propagating events generated aloft. They occurred during the night in a stably stratified boundary layer, which points to gravity waves as the most probable generating mechanism. Such top-down propagating events are ubiquitous in the stable atmospheric boundary layer (Einaudi *et al.*, 1989).

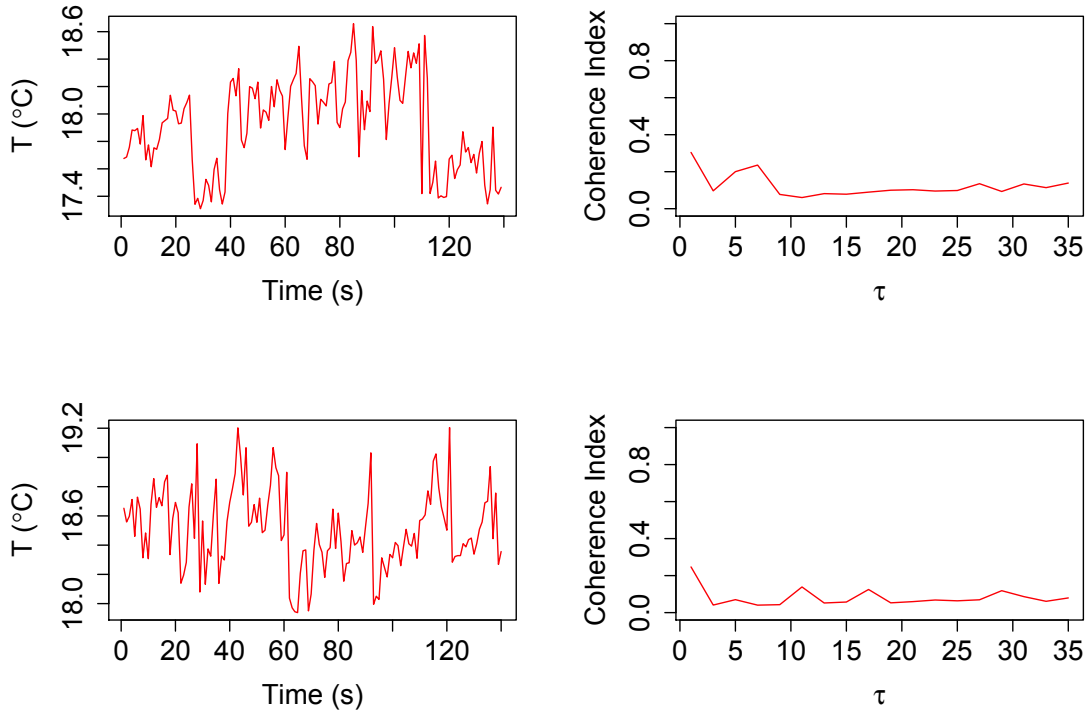

Figure 3.2: *(Color online) Two coherent structure examples from the CASES-99 temperature data (T), and their coherence index as a function of time lag* τ*. The maximum* τ *is* 1/4 *of the time series length (see Section 3.2.3). The maximum* nm_{npe} *values for all considered embedding dimensions m (see Section 3.2.4) of the two events are* −0.09 *and* 0.35*, respectively.*

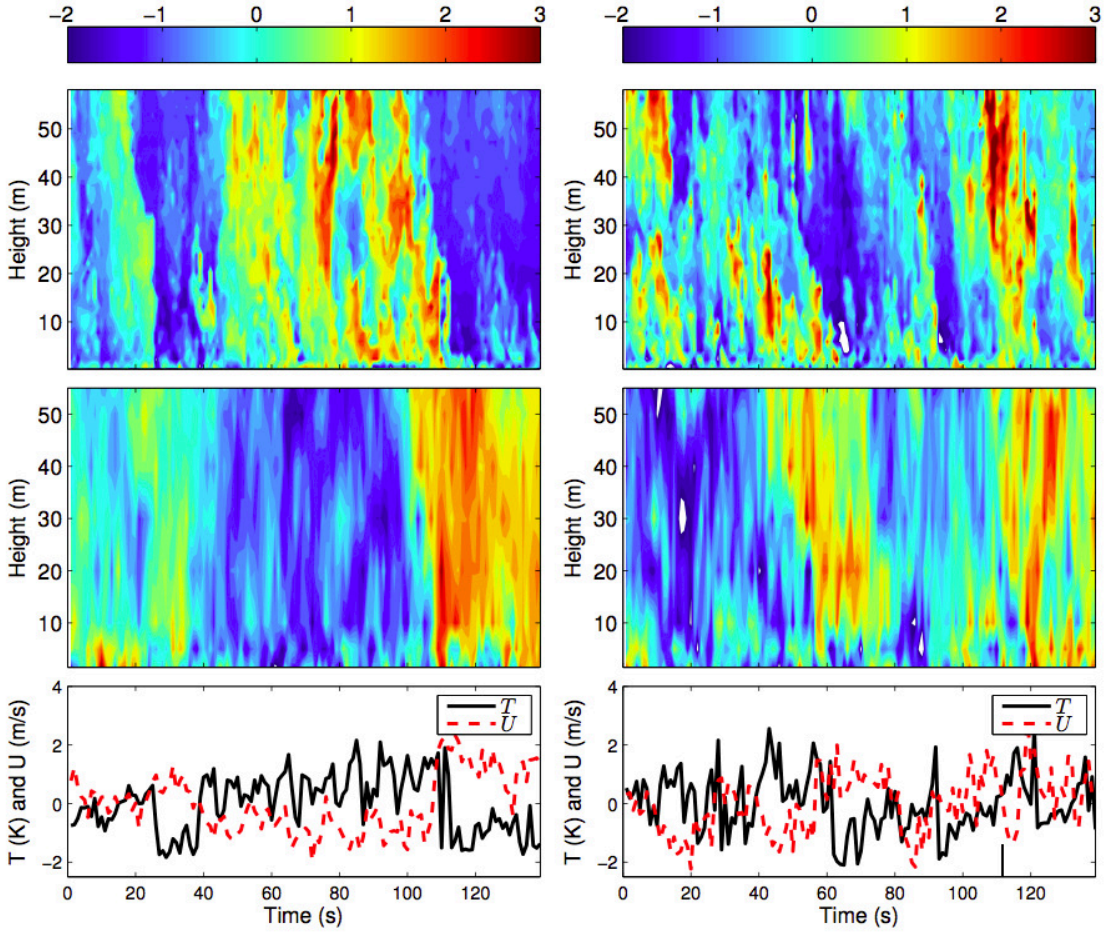

Figure 3.3: *(Color online) Time-height cross-sections of the two coherent structures shown in Fig. 3.2 for (top panels) the normalized temperature perturbation from the 34 thermocouples and (middle panels) the normalized horizontal wind speed from the seven sonics. The bottom panels show time series of the normalized temperature T at 9.5 m and the normalized horizontal wind speed U at the sonic anemometer level 3 (10 m).*

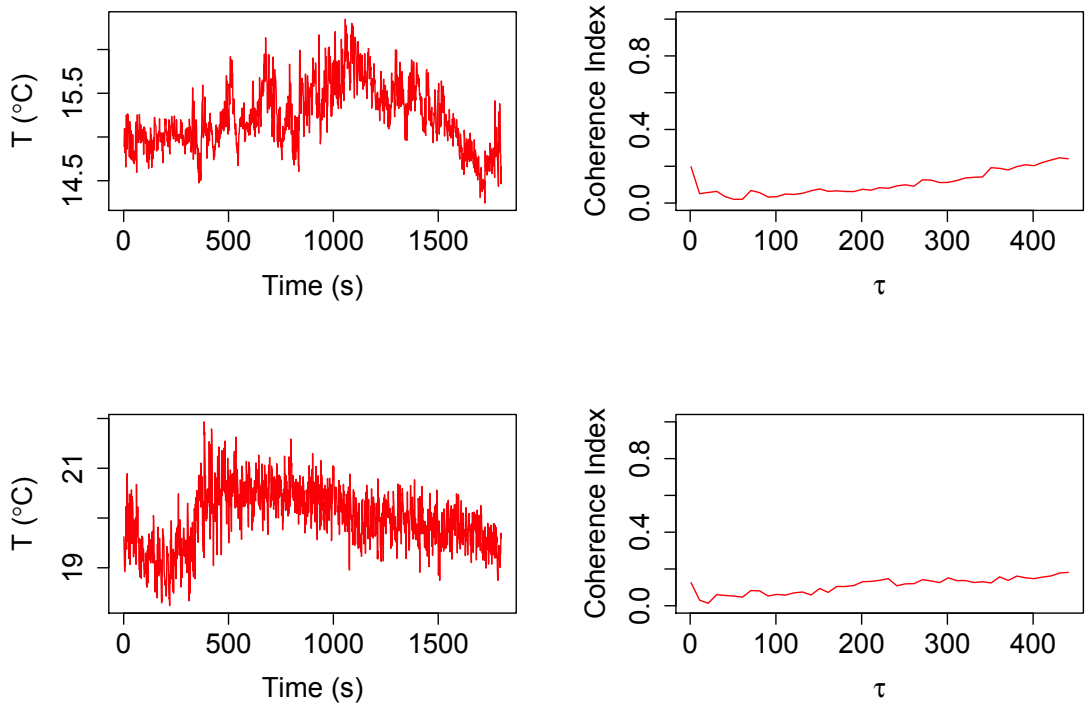

Figure 3.4: *(Color online) As in Fig. 3.2, except for coherent structures at larger scales. The structure onset times are (top panels) 2200 LST 7 October and (bottom panels) 2000 LST 10 October 1999. The maximum* nm_{npe} *values of the two events are* 0.02 *and* 0.53*, respectively.*

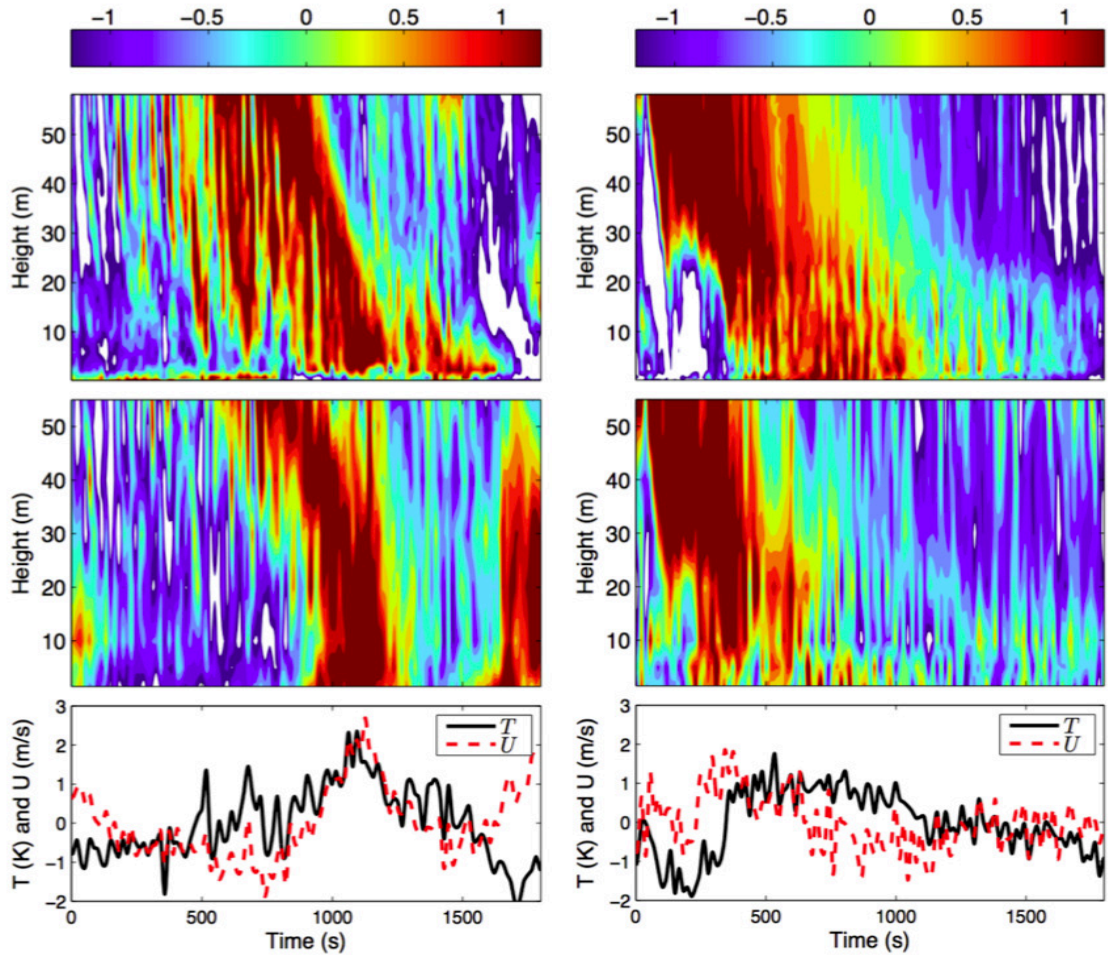

Figure 3.5: *(Color online) As in Fig. 3.3, except that shown are the two coherent structures from Fig. 3.4. For presentation purposes, the two variables are low-pass filtered at 6.2 s using the bi-orthogonal wavelet BIOR5.5.*

3.4 Conclusion

We show that the space and time organized structures in turbulent flow do not necessarily have correlated phases. This warns against assuming that randomizing spectral phases removes all coherent structures from the turbulence time series in all cases. While this assumption is still applicable to many systems as has been shown in previous studies, caution should be used when analyzing complex real-world turbulent flows.

Equivalently, using the term "coherent structure" might not be sufficient to transfer the true meaning without additional description. For example, the phase correlation of coherent structures is not frequently assumed in atmospheric science studies, which in practice may result in a considerable increase in the number of detected events. On the other hand, in cases where coherent structures need to be removed from the data, phase randomization is a typically used tool, but may not be reliable. Any comparison of results of studies using such differently defined coherent structures is not advisable, especially if the goal is to review the common dynamics or effects of coherent structures.

PART B: Suggested Declaration for Thesis Chapter

Monash University

Declaration for Thesis Chapter 4

Declaration by candidate

In the case of Chapter 4, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
- Developed, established and verified the method - Wrote programming codes and the article	90

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Kate Smith-Miles	Provided helpful guidance and editorial work	
Danijel Belusic	Provided helpful guidance and editorial work	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature	[redacted]	**Date** 21/05/2014
Main Supervisor's Signature	[redacted]	**Date** 21/5/14

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Chapter 4

How to Extract Meaningful Shapes from Noisy Time-Series Subsequences?

Chapter 4 is based on the article Kang Y, Smith-Miles K, Belušić D. 2013. How to extract meaningful shapes from noisy time-series subsequences? In:* Proceedings of the 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). *IEEE, pp. 65–72, doi: 10.1109/CIDM.2013.6597219.

Abstract. *A method for extracting and classifying shapes from noisy time series is proposed. The method consists of two steps. The first step is to perform a noise test on each subsequence extracted from the series using a sliding window. All the subsequences recognized as noise are removed from further analysis, and the shapes are extracted from the remaining non- noise subsequences. The second step is to cluster these extracted shapes. Although extracted from subsequences, these shapes form a non-overlapping set of time series subsequences and are hence amenable to meaningful clustering. The method is primarily designed for extracting and classifying shapes from very noisy real-world time series. Tests using artificial data with different levels of white noise and the red noise, and the real-world atmospheric turbulence data naturally characterized by strong red noise show that the method is able to correctly extract and cluster shapes from artificial data and that it has great potential for locating shapes in very noisy real-world time series.*

Keywords. *Shape Extraction; Noisy Time Series; White Noise Test; Red Noise Test; Clustering.*

References are considered at the end of the thesis.

Chapter 4

How to Extract Meaningful Shapes from Noisy Time-Series Subsequences?

4.1 Introduction

Time series mining can be found within a broad range of application domains. It becomes vital with the increasing demand for understanding the underlying processes or phenomena that generate the data. A specific interesting time series mining task is to detect shapes (patterns) in a given noisy time series, which are not noise and have meaningful interpretation in the real world.

Pattern searching in data is a hot topic in a diverse range of application fields. For example, in bioinformatics, much attention has been given to the problem of sequence matching and sequence pattern identification (Guan *et al.*, 1996; Benson and Waterman, 1994; Rigoutsos and Floratos, 1998; Tompa, 1999; Bailey *et al.*, 2009). For real-valued sequences (i.e., time series), numerous algorithms have been

proposed to detect pre-defined patterns (Keogh *et al.*, 2000; Belušić and Mahrt, 2012; Keogh *et al.*, 2001a, 2002; Das *et al.*, 1998; Agrawal *et al.*, 1993; Pong Chan and Fu, 1999; Hochheiser and Shneiderman, 2002; Singh, 2000). However, in most cases, the patterns in time series are not known in advance. Therefore, the extraction of previously unknown, frequently occurring patterns in time series has been recently recognized as an interesting task (Lin *et al.*, 2002; Mueen *et al.*, 2009b; Chiu *et al.*, 2003; Mueen *et al.*, 2009a; Lam *et al.*, 2011). To detect shapes in time series, an initial idea was to cluster all the subsequences extracted using a sliding window and get the prototypes of the typical shapes in the series. However, the sliding window leads to meaningless clustering results as it always forms sine wave cluster centres regardless of the data set (Keogh *et al.*, 2003). To avoid this problem, Keogh *et al.* (2003) proposed a solution by considering the concept of time series motifs. Recently, Lin *et al.* (2002) proposed an efficient algorithm to find time series motifs by using the discrete representation of time series. The exact motif discovery algorithm proposed by Mueen *et al.* (2009b), used as a comparison in this paper in Section 4.4.1, automatically constructs "dictionaries" of recurring patterns in a faster and efficient way. Chiu *et al.* (2003) introduced a time- and space-efficient method to find motifs in time series based on a pattern discovery algorithm in biosequences.

The aim of this paper is to propose a new method for extraction of typical shapes which repeat within time series, especially in very noisy time series data. The method proposed in this paper is also based on sliding windows, but the clustering is performed in a meaningful way. The approach will be introduced in detail in Section 4.2. The main contribution of our approach is that it avoids the meaningless subsequence clustering interpretation in Keogh *et al.* (2003) and finds frequently occurring shapes in a meaningful way. More importantly, the method is applicable to different noise types and levels in the time series, which is very advantagous when dealing with very noisy time series.

The rest of this paper is organised as follows. Section 4.2 presents the new method. Section 4.3 introduces the data used in the experiments. Section 4.4 presents the experiment results for artificial time series and real world time series and compares the results with some related work. Finally in section 4.5 we draw some conclusions and highlight directions for future work.

4.2 Proposed Methodology

The method proposed in this paper consists of two steps. In the first step, a noise test is performed on sliding subsequences extracted from a time series. The shapes are located and detected based on the noise test p values. The details are given in Section 4.2.1 and Section 4.2.2. This step is at the same time the major strength and weakness of the method: the noise test ensures that the method is able to distinguish signal from noise even with high noise levels, but the method is not applicable to time series without noise separating individual shapes. The second step is to group the shapes found in the first step into clusters. The clustering method used here yields members of the same cluster behaving similarly in terms of shape characteristics. Details of the second step will be demonstrated in Section 4.2.3.

4.2.1 Noise Test in Time Series

White Noise Test

White noise is a model mostly frequently seen in time series. We use Ljung-Box test to show whether the data are independently distributed (Box and Pierce, 1970). It is defined as:

H_0: The data are independently distributed.

H_1: The data are not independently distributed.

The test statistic is:

$$Q = n(n+2)\Sigma_{k=1}^{h} \frac{(\hat{\rho_k^2})}{n-k} \tag{4.2.1}$$

where n is the sample size, $\hat{\rho_k}$ is the sample autocorrelation lag k, and h is the number of lags being tested. We set $h \approx \ln(n)$, which is suggested by simulation studies in Tsay (2005).

Red Noise Test

Putting the frequent use of white noise in time series aside, analysis of time series related to many areas, such as climate, relies on 'red noise' as a simple model for correlation in the series (Percival, 2010). That is why we introduce the red noise test as an alternative to the white noise test for certain applications. Equation (4.2.2) shows an autocorrelation model with lag 1 and an error term represented as white noise. Such an autocorrelation model is called red noise, or a first order Markov process (Percival, 2010). Simply speaking, red noise is interpreted as a first-order autoregressive ($AR(1)$) stationary Gaussian process at unit lag (Storch and Zwiers, 1999).

$$x(t) = \phi x(t-1) + \epsilon(t) \tag{4.2.2}$$

where $x(t)$ is the value of variable s at time t, ϕ is the autocorrelation coefficient with lag 1, $\epsilon(t)$ is the value of white noise function at time t.

To test whether a series is red noise, firstly we fit an $AR(1)$ model on the series, and then perform white noise test on the model residuals. If the white noise test shows that the $AR(1)$ model residuals are white noise, then the given series is claimed to be red noise.

4.2.2 The First Step: Shape Extraction in Time Series

To extract shapes, we first perform noise test on each individual subsequence extracted with a sliding window.

Definition 1. *A **subsequence** $x_q(t)$ for a time series $x(t)$ with length m is*

$$x_q(t) = (x(t_q), \cdots, x_(t_{q+w-1})) \tag{4.2.3}$$

for $1 \leq q \leq m - w + 1$, where w is the sliding window size, which is also the length of the subsequence.

The sliding window size w is pre-chosen according to the real world context. The sequential subsequences in Equation (4.2.3) are extracted using a sliding window with a length of w and position q, which is incremented with a natural number $\mathbb{N}$. A noise test is performed on each extracted subsequence. Assuming the test p value of the qth subsequence $x_q(t)$ is p_q, we can obtain a p value series for the time series $x(t) : p_1, p_2, \cdots, p_{m-w+1}$. We define a subsequence as a shape if its noise test p value is smaller than a predefined significant level. If there is a consecutive sequence of subsequences defined as shapes according to the noise test, we pick the middle one to avoid fractional shapes which don't contain a complete pattern. The formal definitions are as follows.

Definition 2. ***A shape** is a subsequence whose noise test p value is smaller than a predefined significant level α ($\alpha = 0.05$ is used in this paper).*

Definition 3. *Assume there exists a consecutive sequence of p values $p_s, p_{s+1}, \cdots, p_t$ which satisfies: (1) $p_i \leq \alpha, i = s, s+1, \cdots, t$; (2) $t - s \geq w/2$, then we define the subsequence $x_{\lfloor \frac{t+s}{2} \rfloor}(t)$ as **the shape** we are interested in.*

The first step assumes the existence of noise regions that separate individual shapes. This is a crucial assumption and limitation of this method, because

otherwise it could not distinguish between different shapes. As we also see, this step extracts the shapes from the time series without organising them in categories or clusters. Therefore, the following second step is required in order to cluster the extracted shapes.

4.2.3 The Second Step: Clustering of the Extracted Shapes

Since the shapes extracted in the first step are not overlapping, any clustering algorithm could be used for their classification. In order to account for the global characteristics of the shapes, a feature-based Ward's hierarchical clustering method will be used here (Wang *et al.*, 2006). It clusters the extracted shapes using the Euclidean distances among a set of features calculated from the raw data, rather than the distances among the raw data itself. In this paper, the following features of subsequences are used: standard deviation, non-linearity (Wang *et al.*, 2006), serial-correlation (Wang *et al.*, 2006), trend, maximum, minimum, standard deviation and serial-correlation of the first order difference of the subsequences. The feature set can be chosen for a specific application to best capture the underlying characteristics of the shapes. In a word, the second step groups the n shapes extracted from the first step in an d-dimension (with $d = 8$ using the mentioned feature set) feature space to obtain the k typical shapes extracted from noise.

4.3 Experimental Data

Here, we demonstrate how the artificial time series were generated and introduce the real world data used in the experiments.

4.3.1 Artificial Time Series

We will use the shapes in the classic Cylinder-Bell-Funnel data (Keogh and Kasetty, 2002). This dataset consists of random instantiations of the patterns, with Gaussian noise added. Besides these three shapes, we will also include a single-cycle sine function as another shape supposed to be extracted. Figure 4.1 shows an instance of each of the four patterns.

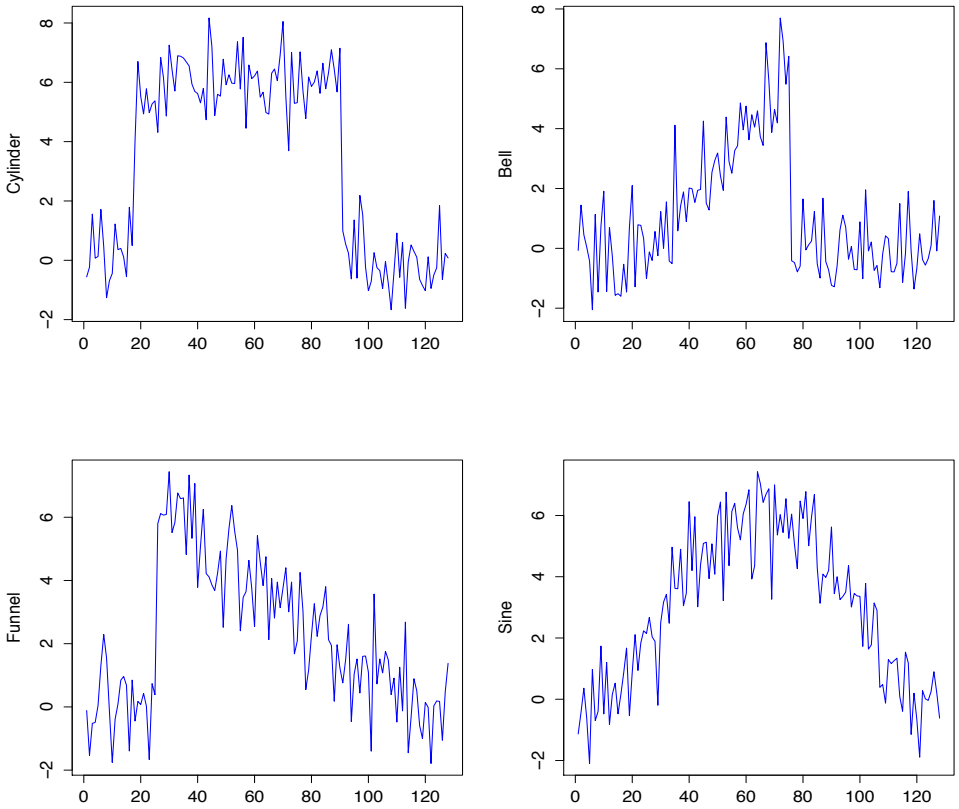

Figure 4.1: *Examples of Cylinder, Bell, Funnel and sine shapes.*

Artificial Time Series with White Noise

Based on the four shapes shown above, we generated a dataset that contains 5 instances of each shape, which are randomly concatenated with each two neighbouring instances separated by a white noise time series with the same length,

which is 128. The white noise is generated using a random number generator that follows a normal distribution $N(0, \sigma^2)$. The generated time series with $\sigma = 1$ is shown in the top panel of Figure 4.2.

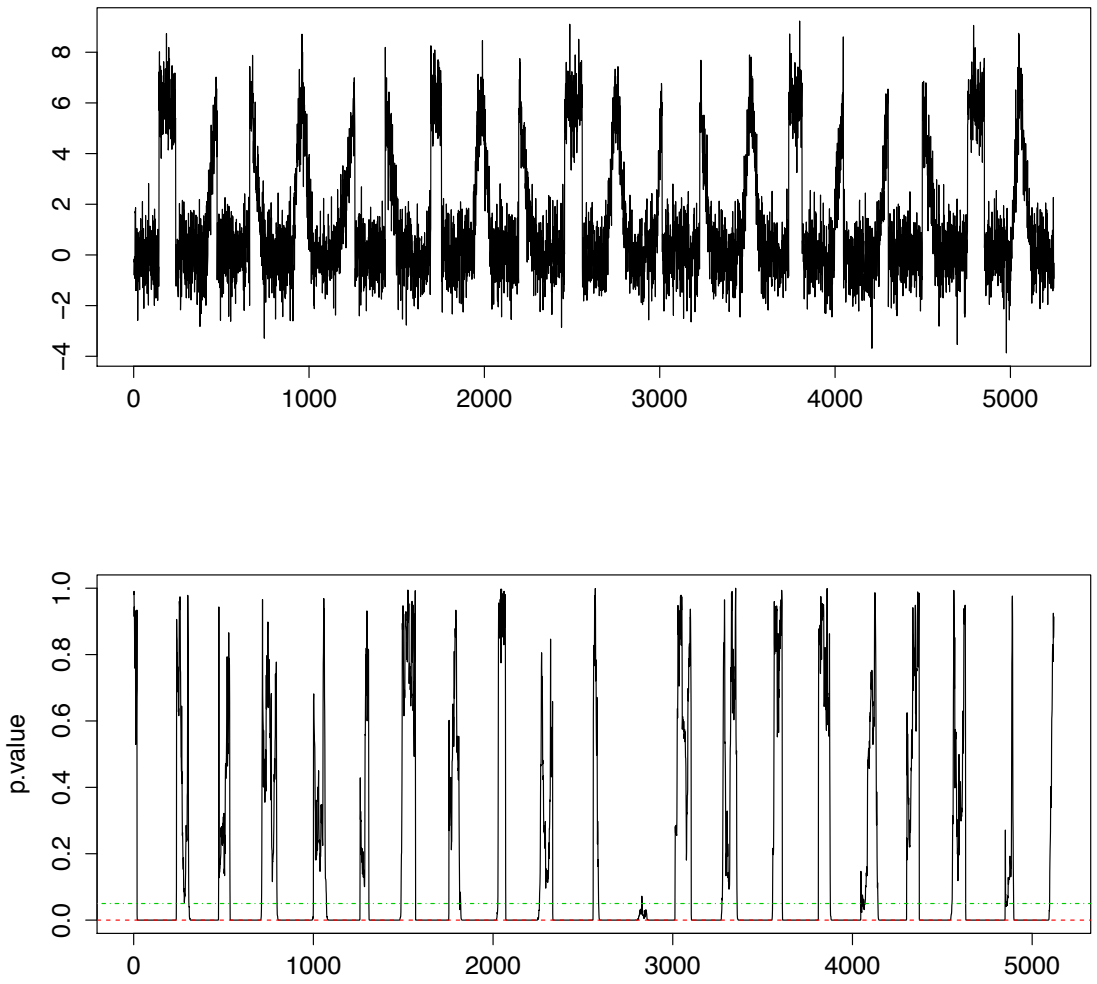

Figure 4.2: *Artificial time series with white noise with $\sigma = 1$ (top panel) and Ljung-Box test p values for subsequences extracted from the artificial time series (bottom panel). The red dashed and green dotdash lines represent zero line and the threshold $\alpha = 0.05$, which also apply to the following figures.*

Artificial Time Series with Higher Level White Noise

To demonstrate how the method is influenced by the white noise level, we generate time series with higher white noise value, $2 * \sigma$ and $3 * \sigma$ respectively, and obtain

more noisy artificial time series. That is to say, instead of $N(0,1)$, the white noise used to generate artificial time series in Section 4.3.1 follow the distribution $N(0,2^2)$ and $N(0,3^2)$ respectively.

Artificial Time Series with Red Noise

Instead of using white noise, we also generate the time series by including red noise:

$$x(t) = 0.3 * x(t-1) + \epsilon(t), \text{where } \epsilon(t) \sim N(0,1)$$

in the artificial time series.

4.3.2 Real World Time Series: The Temperature and Vertical Wind Speed

The temperature and vertical wind speed turbulence data were measured by a sonic anemometer with 60 Hz sampling frequency. The anemometer was located at the height of 34 m above the ground during the FLOSSslowromancapii@ experiment in northern Colorado (Mahrt, 2011b). Here we average a sample of the measured data over 180 points, which creates the time series with 3s interval as the input to the method (see the top and middle panels of Figure 4.3). The time series are clearly very noisy and it is very difficult to locate interesting non-noise shapes using our eyes. We demonstrate the potential of the new method in finding shapes in these time series in Section 4.4.2.

4.4 Results

4.4.1 Shape Extraction in Artificial Time Series

Artificial Time Series with White Noise

As we know the length of embedded shapes is 128, we use $w = 128$ in this artificial time series to extract shapes. In real world applications, w is determined by experience and robustness. The bottom panel of Figure 4.2 depicts the p value series of the sliding subsequences of length 128 extracted from the artificially generated time series (See the top panel of Figure 4.2). Note here although we know the types of shapes embedded, we assume they are not known in advance, which is mostly the case in reality. Using the proposed algorithm, the shapes are found according to Definition 3, and are shown in Figure 4.4. As can be seen from Figure 4.4, there are exactly the four patterns that were used to generate the time series: cylinder, bell, funnel and sine shapes.

After extracting the shapes, the hierarchical clustering is performed in order to find similar patterns among the shapes. This makes the expected results of the method nominally similar to the mentioned motifs in Lin *et al.* (2002), Keogh *et al.* (2003) and Mueen *et al.* (2009b). Figure 4.5 shows the dendrogram from hierarchical clustering on the extracted shapes based on the features (Wang *et al.*, 2006). The four patterns which form the artificial time series are clearly found from this dendrogram. For the purpose of comparison, Figure 4.6 shows the dendrogram from hierarchical clustering of the shapes based on raw data, without using the features. As we can see from the figure, the four patterns are not distinguished in expected groups.

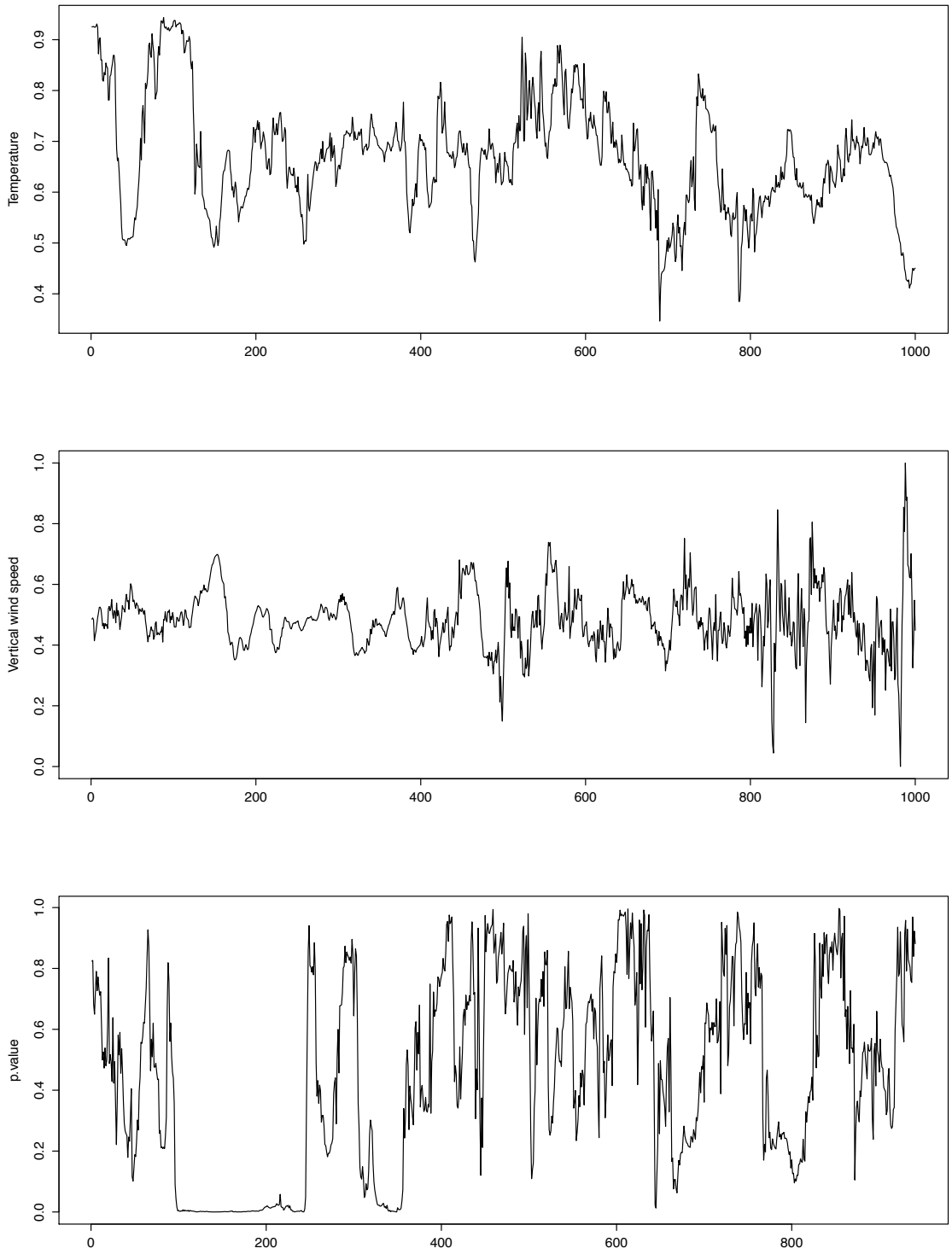

Figure 4.3: *Temperature (top panel) and vertical wind speed time series (middle panel) with 1000 data points measured at the same location; red noise test p values on vertical wind speed time series (bottom panel).*

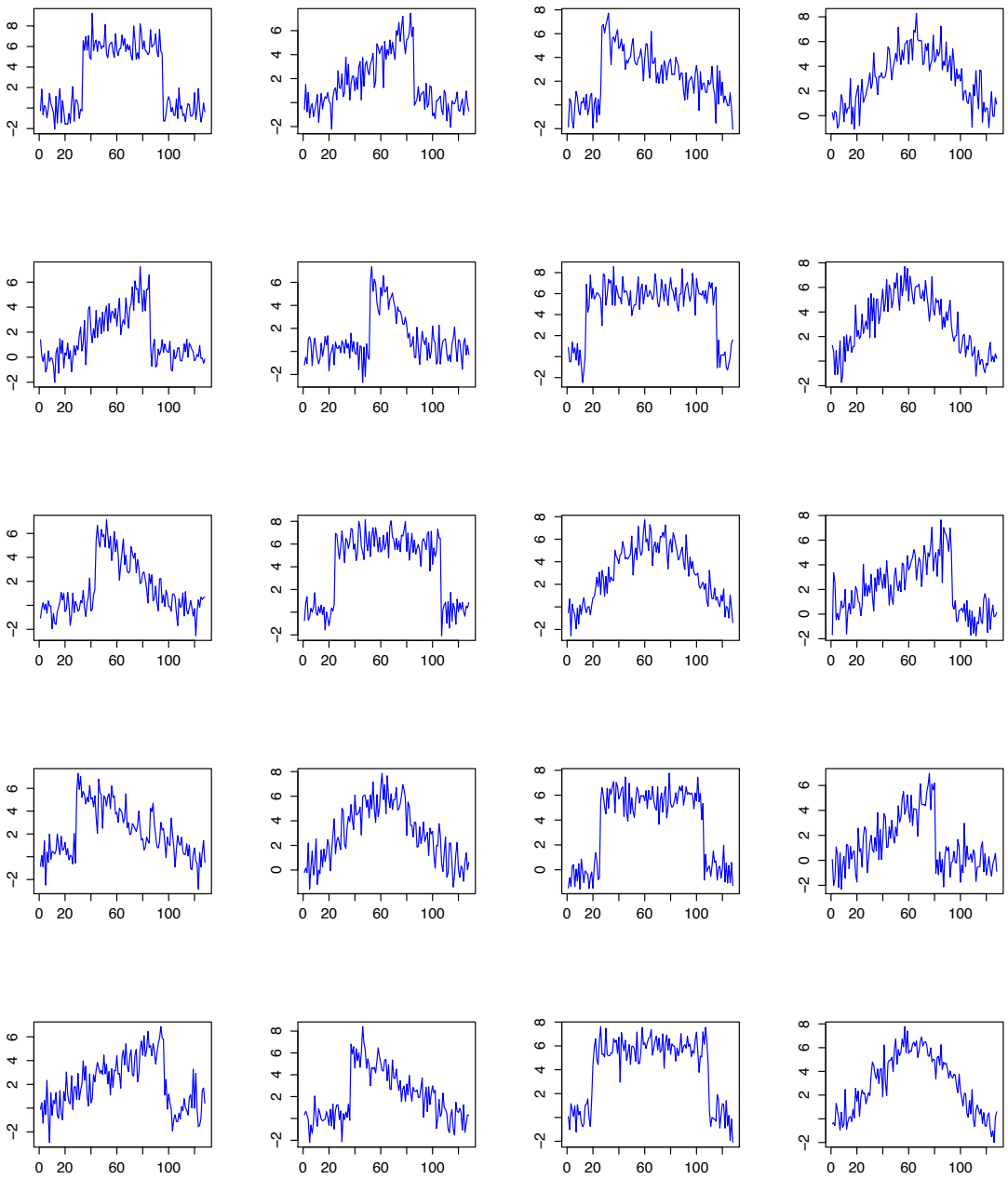

Figure 4.4: *The* 20 *shapes extracted from the artificial time series shown in the top panel of Figure 4.2.*

Furthermore, the clustering can also be performed in an on-line manner. Growing Feature Quantization (GFQ) (Kang, 2012), using only one user defined threshold to control the growth rate of the cluster formation, is able to do on-line clustering on the extracted shapes. As soon as the shape is detected, the GFQ method decides automatically which cluster this shape belongs to. Figure 4.7 gives the relationship

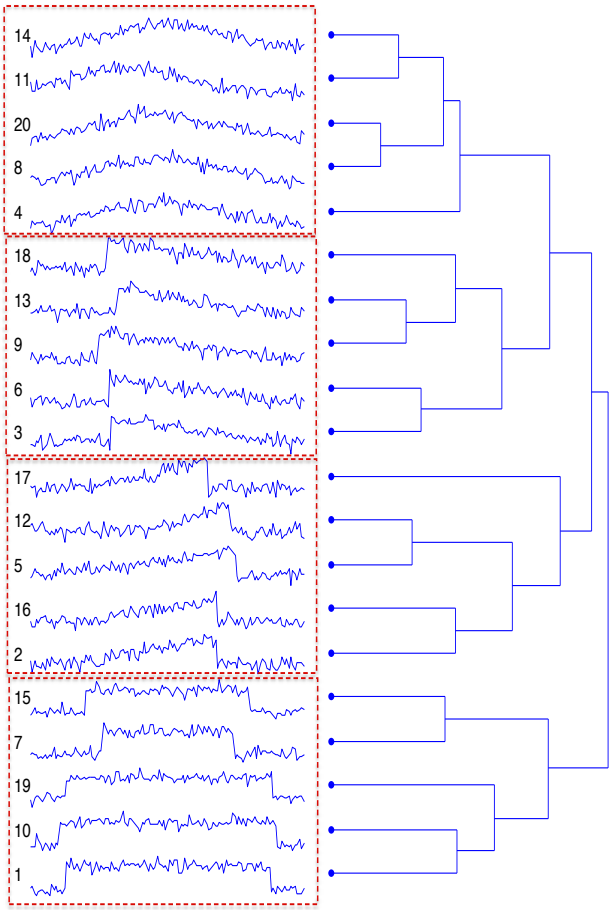

Figure 4.5: *Dendrogram from hierarchical clustering of the extracted shapes based on features.*

between the threshold used in GFQ and the number of clusters obtained. The figure suggest that there are four clusters because this is associated with the largest range of threshold values. The results of clustering shapes using GFQ are shown in Figure 4.8. Each colour of shapes represents a single cluster category. Although the fourth shape is mis-clustered, it is still thought provoking since the clustering is performed in an real-time manner. And the reason for the mis-clustering might be because the fourth shape is still in the early stage of learning process of GFQ.

In order to compare our results with the motif discovery algorithm in Mueen *et al.* (2009b), we use their algorithm on the same artificial data by setting the factor of the cluster radius X to be $X = 1.5$ and number of clusters $K = 6$ to show the first 6 motifs. Figure 4.9 shows the motifs of length 128 obtained by their algorithm

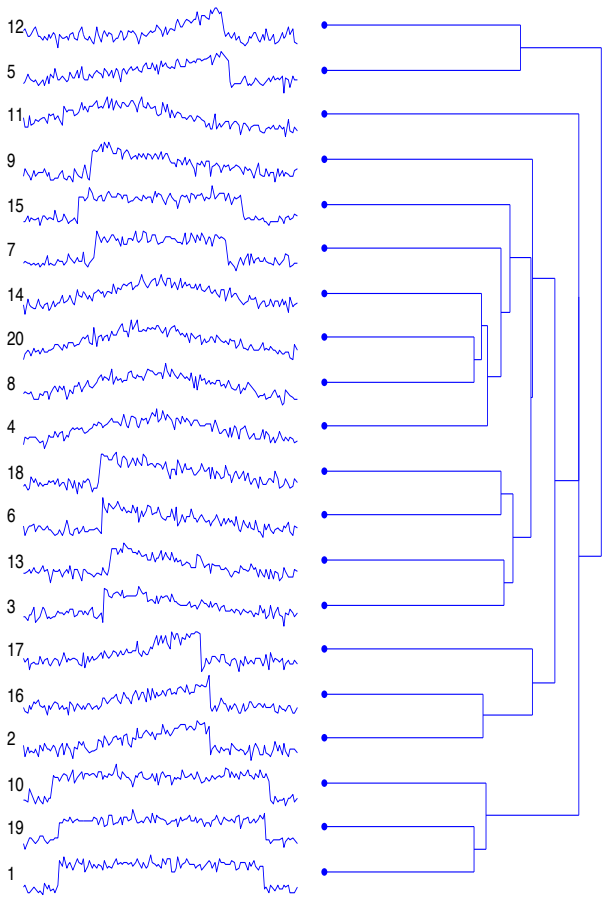

Figure 4.6: *Dendrogram from hierarchical clustering of the extracted shapes based on raw data.*

in the entire time series shown in the top panel of Figure 4.2. We can see that in the first two motifs, there are different patterns mixed together in the same cluster. In order to avoid this, we try to decrease the factor of the cluster radius to be $X = 1.3$. Figure 4.10 shows the motifs obtained in this case. The algorithm is automatic and efficient, and the motifs found using Euclidean distance are quite meaningful. However, because of the existence of white noise among shapes, some motifs found, e.g., the members in the third and fourth motif in Figure 4.9 have plentiful noise included and those in the fifth motif are actually white noise. We can also see that in Figure 4.10, there are a number of subsequences which are noise being identified as a member of motifs. The approach proposed here can address this problem by only considering the non-noise subsequences in time

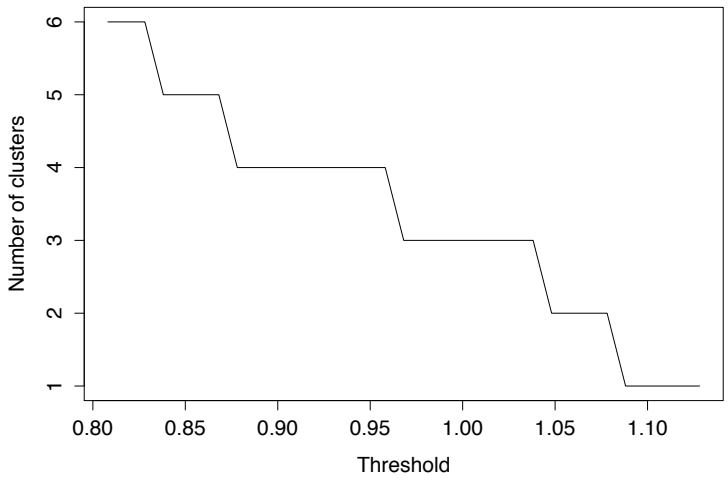

Figure 4.7: *The relationship between the threshold used in GFQ and the number of clusters obtained.*

series to be shapes. Furthermore, the new approach clusters the shapes based on features instead of euclidean distances among raw subsequences. This keeps the global shape (motif) characteristics and avoids subsequences which are similar in euclidean distances but distinct in shapes from being grouped together. On the other hand, it helps subsequences that are similar in shapes but with different shifts or lengths, e.g., the Cylinder shape members in Figure 4.5, being grouped in the same cluster.

Artificial Time Series with Higher White Noise Levels

From Section 4.4.1, the pre-embedded shapes are well recognised using the new method under a certain noise level. But the problem will become more challenging with the noise level increased. In order to demonstrate that the method is applicable under a high-noise level, we increase the noise levels of the artificial time series in Section 4.4.1 to $2*\sigma$ and $3*\sigma$. The corresponding p values are shown in Figure 4.11 and Figure 4.12, respectively. Even with the increase of the noise level to three times of the original, the method still recognizes the shapes. With

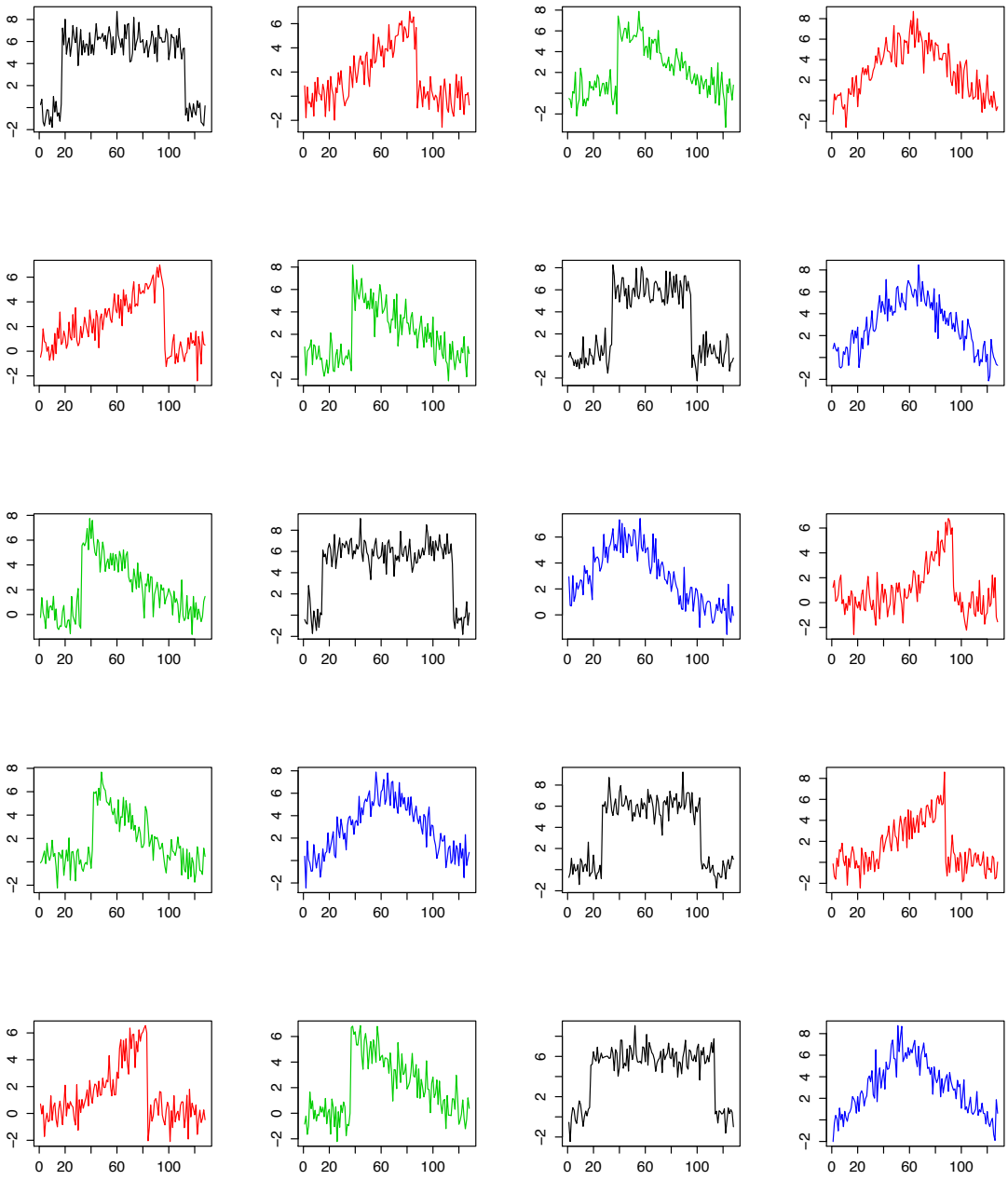

Figure 4.8: *The clustering results of the extracted shapes based on GFQ; four colours represent four different clusters.*

that level of noise, the visual recognition of shapes from the time series would be very difficult.

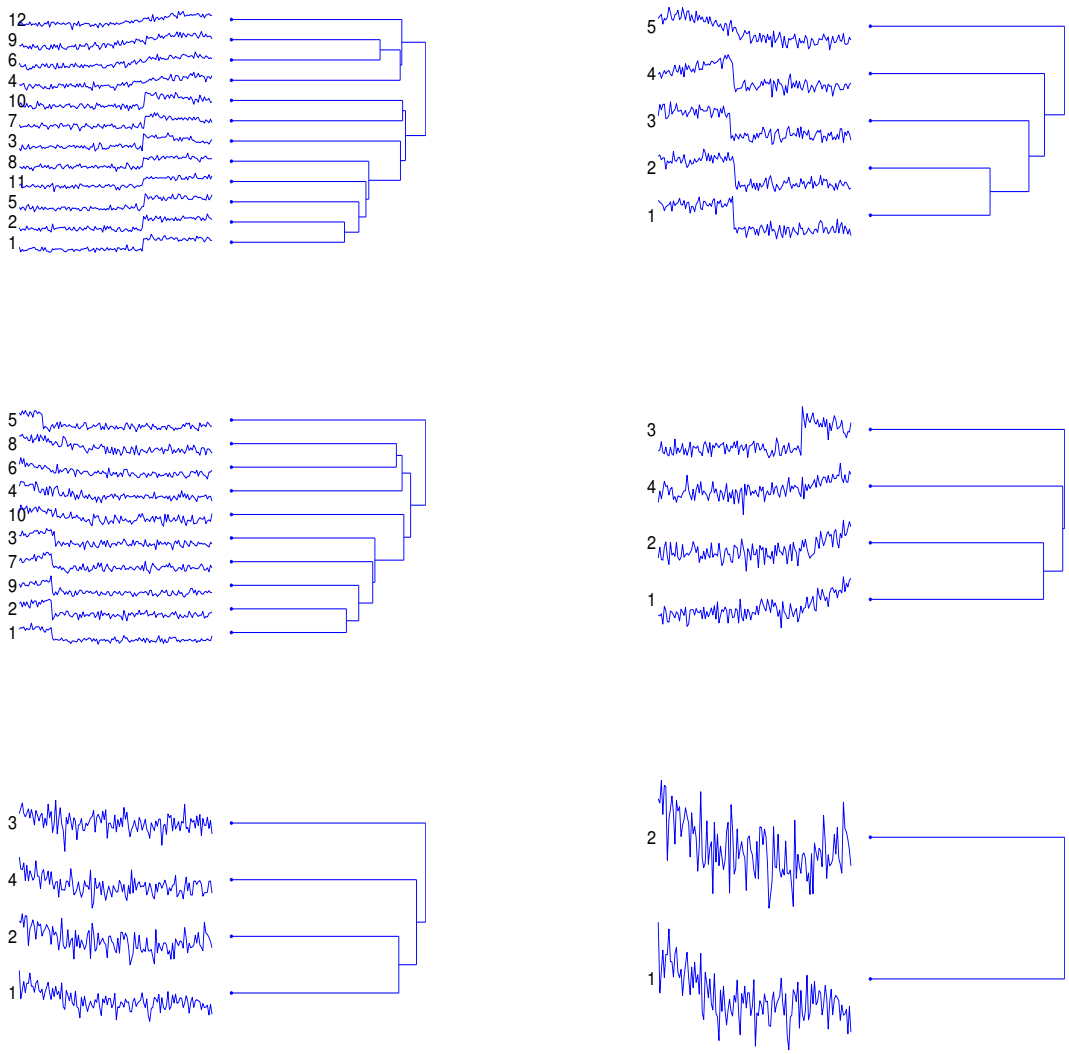

Figure 4.9: *The six motifs found using the algorithm in Mueen* et al. *(2009b) ($X = 1.5$).*

Artificial Time Series with Red Noise

As a step towards real-world data, particularly in broad field of geophysics, artificial time series are created with red noise instead of white noise. The shapes are correctly recognised using the red noise test in the method (See the middle panel of Figure 4.13). However, when we perform white noise test on this time series, the p-value series (See the bottom panel of Figure 4.13) can not clearly separate shapes from noise. This proves that the method can be used for various types of noisy data provided the proper assumptions are made on the nature of noise.

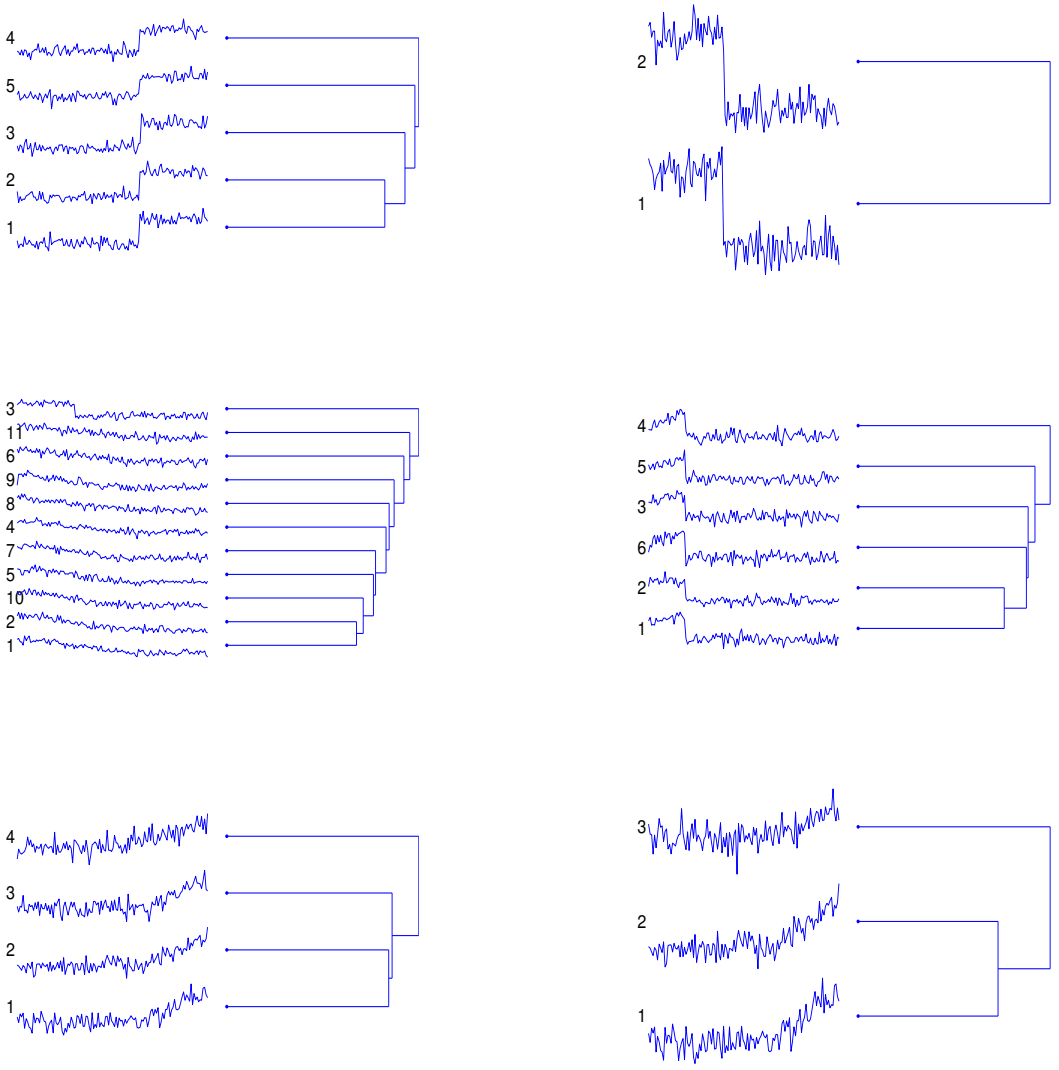

Figure 4.10: *The six motifs found using the algorithm in Mueen* et al. *(2009b) (X = 1.3).*

4.4.2 Real World Application: The Temperature and Vertical Wind Speed Time Series

Using sliding window size $w = 60$, which corresponds to a time scale of 3 minutes, the results of the red noise test on subsequences extracted from the vertical wind speed time series are shown in the bottom panel of Figure 4.3. The figure shows that there is only one segment in which the p values are consecutively smaller than α. Assume p_s is the first subsequence satisfying $p < \alpha$ and p_t is the last one

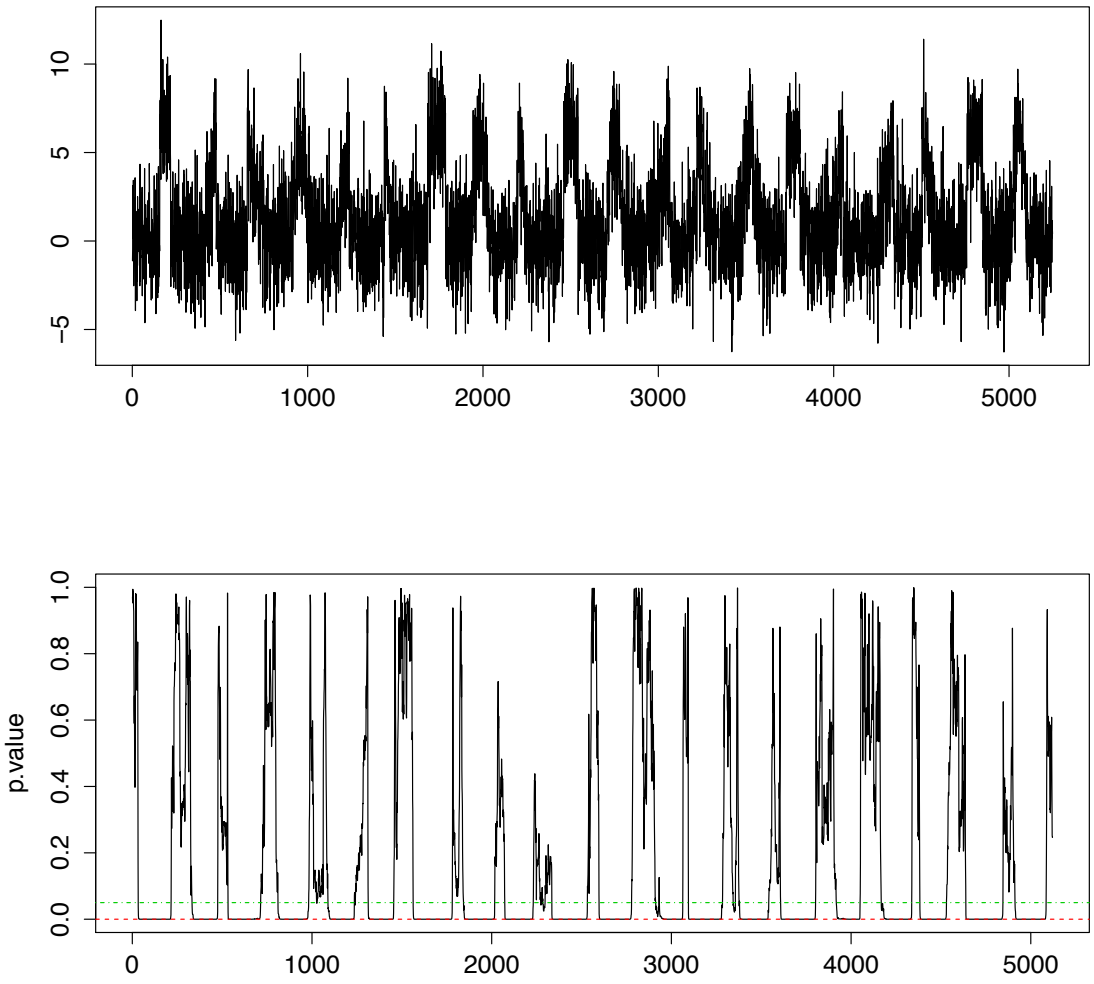

Figure 4.11: *Artificial time series with white noise (top panel) and Ljung-Box test p values for subsequences (noise level is $2 * \sigma$) (bottom panel).*

in this segment meeting $p > \alpha$. In this complicated real world time series, to avoid missing any possible shapes, instead of choosing the middle subsequence as in Definition 3, we choose the entire segment from time point p_s to time point $p_t + w - 1 = p_t + 59$ as the shape. This segment is shown in Figure 4.14, together with the corresponding segment of the temperature time series. The existence of the phase angle between the vertical wind speed and temperature implies a gravity-wave origin of the shape. The visually determined value of approximately $\pi/2$ indicates that this is a signature of an atmospheric internal gravity wave

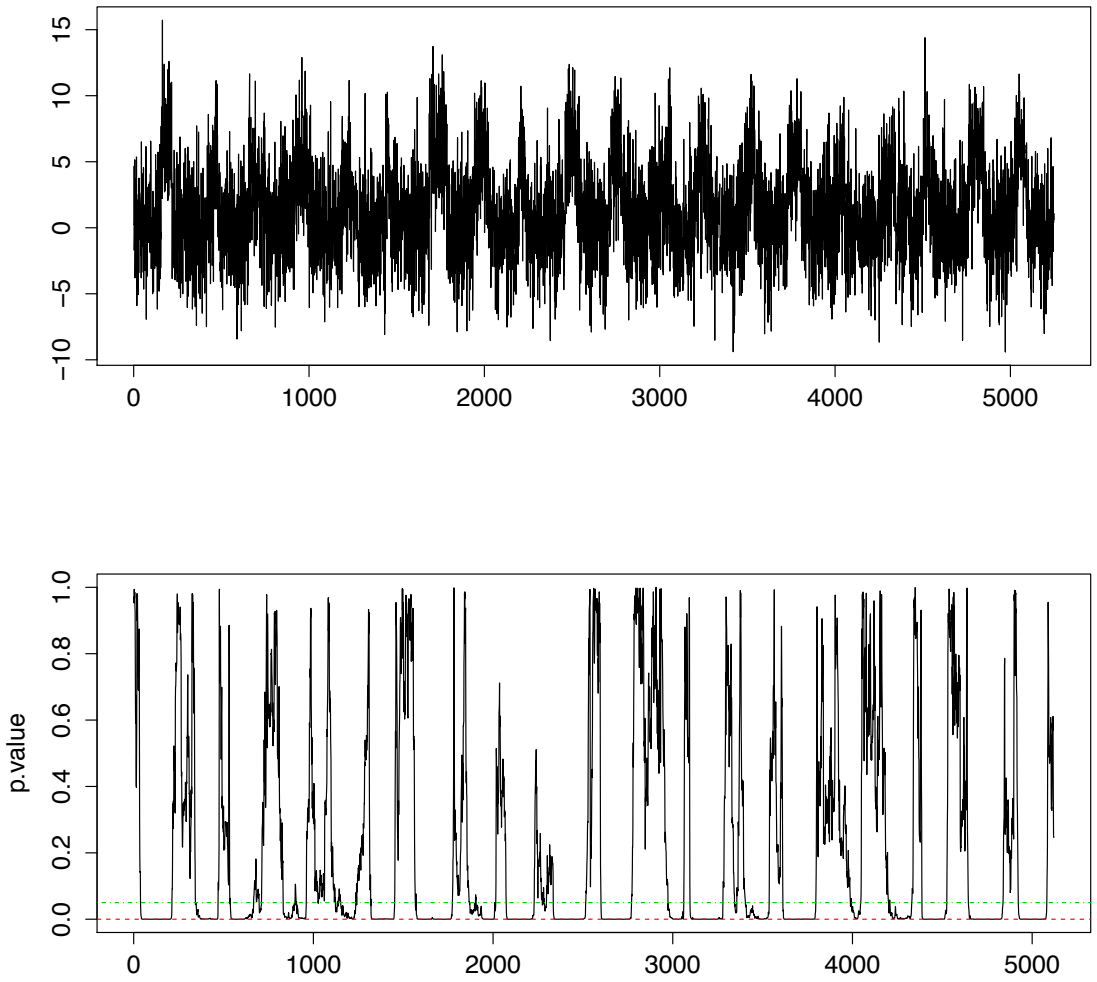

Figure 4.12: *Artificial time series with white noise (top panel) and Ljung-Box test p values for subsequences (noise level is* $3 * \sigma$*) (bottom panel).*

(Belušić and Mahrt, 2012). Further detailed analysis would be required to describe the dynamics of this event, but this is beyond the scope of this study. However, this example illustrates that the method is capable of extracting physically meaningful shapes from real-world turbulence time series, which are among the most complex time series found in nature.

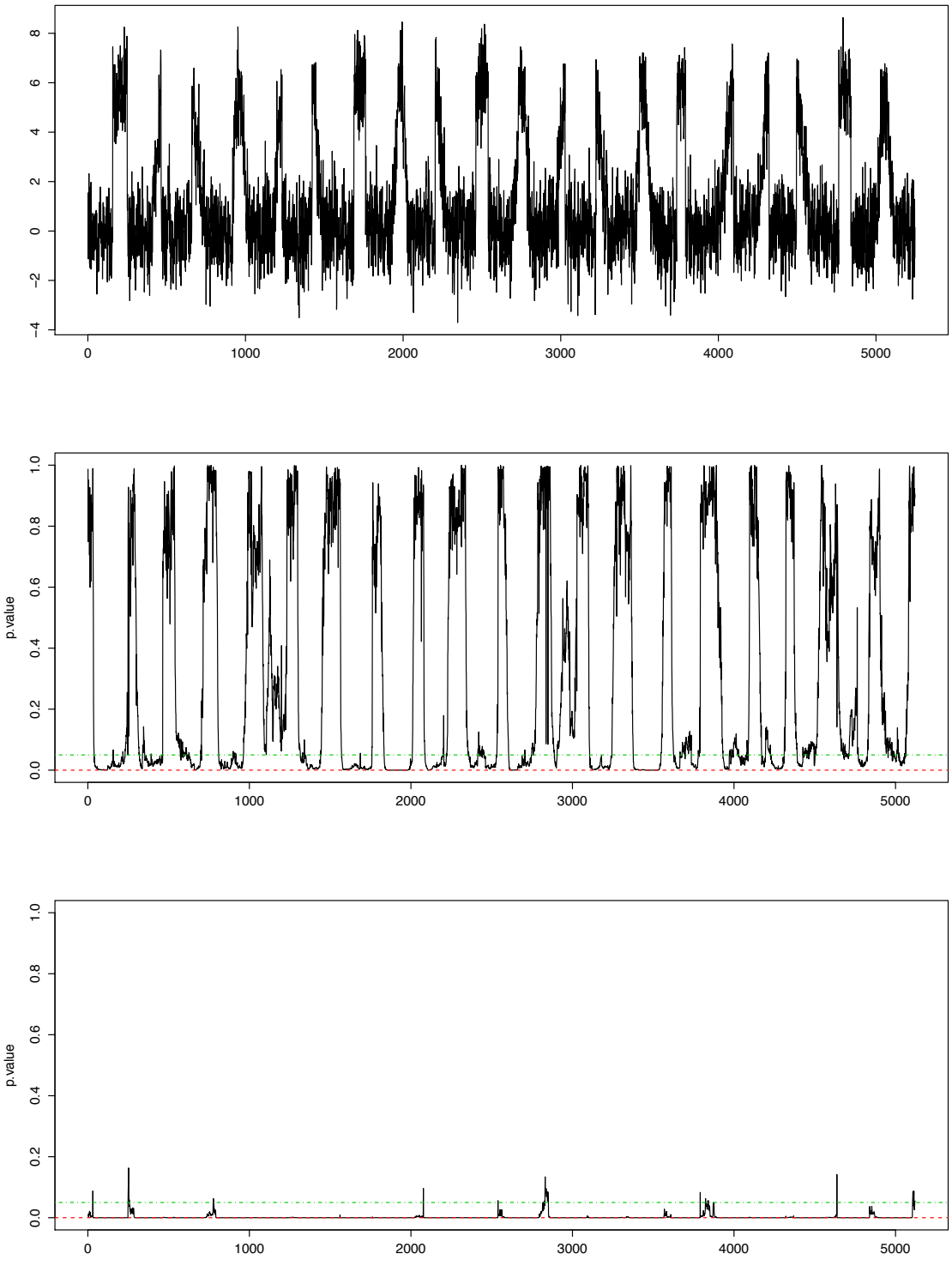

Figure 4.13: *Artificial time series with red noise (top panel), red noise test p values for subsequences (middle panel) and white noise test p values for subsequences (bottom panel).*

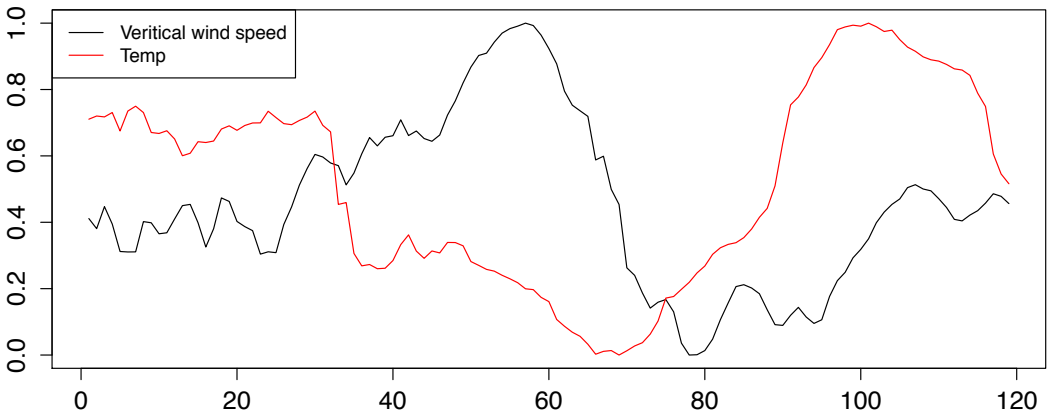

Figure 4.14: *The detected shapes from the vertical wind speed and temperature time series.*

4.5 Conclusion

A new method for shape extraction from time series is proposed. It is based on two steps: a noise test, which is performed on each subsequence extracted from the time series, and clustering of the extracted shapes into similar patterns. The second step is based on a set of features, which keeps the information of the main characteristics of shapes and is shown to yield better results than the clustering based on raw data. Shape patterns found using these two steps are compared with the motif discovery algorithm proposed in Mueen *et al.* (2009b). The main contribution of the new method is that it ignores the noise part in time series and focuses on the non-noise subsequences, which improves the meaningfulness of the shape searching procedure. Furthermore, the proposed method is robust to higher noise levels, which is a strong advantage regarding very noisy time series.

The proposed method is applied to both artificial data and real world data. Shapes used to generate the artificial data are exactly found using the method. Regarding the real world time series, the results show that the method has potential for

application. More research is necessary to further study the real world time series and investigate the extraction and interpretation of shapes.

PART B: Suggested Declaration for Thesis Chapter

Monash University

Declaration for Thesis Chapter 5

Declaration by candidate

In the case of Chapter 5, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
- Developed, established and verified the method - Wrote programming codes and the article	90

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Danijel Belusic	Provided helpful guidance and editorial work	
Kate Smith-Miles	Provided helpful guidance and proofreading	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature	[redacted]	**Date** 21/05/2014
Main Supervisor's Signature	[redacted]	**Date** 21/5/14

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Chapter 5

Detecting and Classifying Events in Noisy Time Series

Chapter 5 is based on the article Kang Y, Belušić D, Smith-Miles K. 2014c. Detecting and classifying events in noisy time series.* Journal of the Atmospheric Sciences *71(3): 1090–1104, doi: 10.1175/JAS-D-13-0182.1.

Abstract. *Time series are characterized by a myriad of different shapes and structures. A number of events that appear in atmospheric time series result from as yet unidentified physical mechanisms. This is particularly the case for stable boundary layers, where the usual statistical turbulence approaches do not work well and increasing evidence relates the bulk of their dynamics to generally unknown individual events.*

This study explores the possibility of extracting and classifying events from time series without previous knowledge of their generating mechanisms. The goal is to group large numbers of events in a useful way that will open a pathway for the detailed study of their characteristics, and help to gain understanding of events with previously unknown origin. A two-step method is developed that extracts events from background fluctuations and groups dynamically similar events into clusters. The method is tested on artificial time series with different levels of complexity and on atmospheric turbulence time series. The results indicate that the method successfully recognizes and classifies various events of unknown origin and even distinguishes different physical characteristics based only on a single-variable time series. The method is simple and highly flexible, and it does not assume any knowledge about the shape geometries, amplitudes, or underlying physical mechanisms. Therefore, with proper modifications, it can be applied to time series from a wider range of research areas.

References are considered at the end of the thesis.

Chapter 5

Detecting and Classifying Events in Noisy Time Series

5.1 Introduction

Time series can be regarded as progressions of various shapes in time. In a broader geophysical context, shapes, or events, are embedded in various levels of noise that is usually of a certain type or color. The motions in the atmosphere exhibit scale interactions such that the power spectra usually decrease with scale as a negative power of the wave number. This is the characteristic shared with red noise, and as a result atmospheric time series are frequently modeled using a first-order autoregressive process ($AR(1)$). Individual studies of atmospheric time series predominantly focus on a relatively narrow range of scales, particularly when describing the underlying dynamical processes. Practically this means that the distinction between noise and "meaningful" features will depend on the scale under consideration. For example, scales of atmospheric waves vary from large planetary waves to those limited by the atmospheric stability at the small end. Researchers interested in planetary waves will tend to disregard

small-scale atmospheric boundary layer (ABL) waves and other processes as noise. Likewise, at smaller, turbulence scales, distinct fluctuation patterns frequently occurring in turbulent flows are termed coherent structures. Coherent structures are distinguished from background fluctuations or noise, and are examined with the goal of understanding important physical characteristics of turbulent flows in terms of elementary structures (e.g., Chen and Hu, 2003; Thomas and Foken, 2005, 2007b; Barthlott *et al.*, 2007).

The usual approach for studying various structures in atmospheric time series is to assume that a certain familiar physical process results in a specific recognizable temporal trace, and to then search for such a trace in the time series. This can be accomplished by searching for certain geometries, such as sine functions for waves and ramp-cliff patterns for coherent structures, or for certain properties, such as large amplitudes or sharp changes (e.g., Antonia *et al.*, 1979; Wilczak, 1984; Chen *et al.*, 1997; Barthlott *et al.*, 2007; Belušić and Mahrt, 2012; Shapland *et al.*, 2012a,b; Segalini and Alfredsson, 2012). Recent studies of the stable weak-wind ABL indicate that many of the processes that are responsible for the variability in the time series are unknown (e.g., Mahrt, 2011b). Such situations do not allow for the above-mentioned approach, but request the opposite strategy — extracting "meaningful", but unknown events from the time series and then understanding their underlying physical mechanisms.

In the stable ABL, gravity waves, transient drainage flows and other systems occur seemingly randomly and either superimpose on the turbulence or affect it by increasing or decreasing its intensity. Currently there are no general ways to clearly distinguish turbulence from waves and other mechanisms, despite the frequent usage of several pragmatic techniques for that purpose. The usual way to study ABL turbulence is deductive, where the hypothesis of turbulence similarity is used to indirectly infer the characteristics of structures in the flow field. This is achieved by assuming that the turbulence statistical effects, which result from a

myriad of interactions of individual flow structures, are uniquely determined by larger-scale flow and surface characteristics. Here the larger scales are assumed to be separated from turbulence scales, and also to be stationary, homogeneous, and known with sufficient accuracy. While this approach has led to a useful quantification of ABL effects in numerical models, its limitations are becoming increasingly apparent (e.g., Nappo *et al.*, 2014). Another way of studying and improving the understanding of the stable ABL dynamics is inductive. This approach involves analysis and understanding of individual structures found in data, with the potential of generalizing the results provided a significant number of structures could be explained or characterized by such approach. Since the representation of stable boundary layers in atmospheric models is in critical need of improvement (e.g., Baklanov *et al.*, 2011; Holtslag *et al.*, 2013; Nappo *et al.*, 2014) and depends a lot on the understanding of the underlying processes, the inductive approach might contribute to that end.

This study presents a step in that direction. A method is developed for extracting and classifying events in time series automatically, without any pre-assumed or pre-defined characteristics of events in terms of their magnitude, geometry or periodicity. The goal is to recognize and classify different events in order to alleviate further analysis of their behavior and underlying mechanisms. The method is presented and validated against a well-known dataset to ascertain that it can be used for further research. While the primary motivation for developing the method is the study of various structures in the stable boundary layer, the method is not limited by atmospheric stability or the scale of the phenomena.

The paper is organized as follows. The details of the two-step method are discussed in section 5.2. The method is first tested on artificially generated time series with different complexities of noise — white and red noise — hence progressing towards the real-world atmospheric conditions. This is detailed in section 5.3. The method is then applied to atmospheric turbulence data, as discussed in section 5.4.

The basic assumptions of the method are further tested in section 5.5, and the conclusions are drawn in section 5.6.

5.2 Methodology

A number of clustering techniques have been developed and used over the last several decades for classifying structures found in different datasets, including various areas of the atmospheric science (e.g., Weber and Kaufmann, 1995; Elsner, 2003; Pope *et al.*, 2009; Belušić *et al.*, 2013). However, it has been shown that the usual clustering techniques return meaningless results when directly applied to sliding and overlapping time series subsequences, because they always yield cluster centers in the form of a sinusoid, regardless of the dataset (Keogh *et al.*, 2003). Therefore, if the goal of analysis is to extract and classify events from time series, a solution is to employ a preprocessing step before clustering. As a result, the method developed here consists of two steps. The first step extracts events from time series using a simple distinction between signal (i.e., events) and noise, and the second step classifies the events using hierarchical clustering. When distinguishing between events and noise in the first step, the characteristics (i.e., color) of noise are assumed to be known a priori. A specific test for that noise color can then be developed and performed for a given scale of interest. Having performed the noise test, the events are defined simply as those subsequences of time series that are significantly different from the noise.

5.2.1 Noise Tests for Time Series

The first step of the method depends on the specification of the characteristics of background noise in a time series. Here we use two different noise models, which

do not exhaust all possibilities for formulation of noise characteristics in various applications.

White Noise Test

White noise is a process most frequently seen in time series, in which data points at different times are not correlated. The Ljung-Box test is applied here for examining whether data points are independently distributed (Box and Pierce, 1970). The test is defined as:

H_0: Data are independently distributed.

H_1: Data are not independently distributed.

The test statistic is:

$$Q = n(n+2)\Sigma_{k=1}^{h} \frac{(\hat{\rho_k^2})}{n-k}$$

where n is the sample size, $\hat{\rho_k}$ is the sample auto-correlation at lag k, and h is the number of lags being tested. As suggested by simulation studies in Tsay (2005), we use $h \approx \ln(n)$. Under the null hypothesis H_0, the asymptotic distribution of Q is χ^2 with h degrees of freedom. In order to determine whether H_0 should be rejected or not, the probability p of obtaining a test statistic at least as extreme as the actually observed statistic under $\chi^2(h)$ is used. The null hypothesis is rejected when the p value is less than a predetermined significance level α, which is often 0.05, indicating that the observed result would be highly unlikely under the null hypothesis. In our case, this means that the data is not white noise.

Red Noise Test

Red noise is modeled as a first-order autoregressive process $AR(1)$. Given a time series with red noise, the white noise test from section 5.2.1 would not recognize

any part of the time series as noise, hence a separate test for red noise needs to be introduced.

Equation (5.2.1) defines $AR(1)$ as a first-order auto-correlation model with the error term represented by a white noise process:

$$x(t) = \phi * x(t-1) + \epsilon(t) \tag{5.2.1}$$

where $x(t)$ is a time series, ϕ is the first-order auto-correlation coefficient ($0 < \phi < 1$) and $\epsilon(t)$ is the white noise process with standard deviation σ_ϵ. In short, a red noise process can be interpreted as an $AR(1)$ process with positive correlation at unit lag (Storch and Zwiers, 1999; Chen *et al.*, 2013).

Considering that $AR(1)$ modeling is only applicable and limited to stationary processes and that some time series are non-stationary, a stationarity test is applied firstly to the given time series $x(t)$. A non-parametric test called Phillips-Perron (PP) Unit Root Test (Banerjee *et al.*, 1993; Perron, 1988) is used here, as it does not assume any characteristic structure of the data. This test is for the null hypothesis that $x(t)$ has a unit root, i.e. is non-stationary, against a stationary alternative. The test has been implemented in many statistical softwares such as R package **stats** (R Core Team, 2013) and Matlab Toolbox Econometrics. Further details of the test can be found in Banerjee *et al.* (1993) and Perron (1988). If $x(t)$ is non-stationary according to the test, then $x(t)$ is different from red noise because red noise is a stationary process. In this case we assign $p = 0$, which is consistent with $x(t)$ being different from noise (see section 5.2.1). Otherwise, if $x(t)$ is stationary, the following red noise test, which is based on the definition of an $AR(1)$ process, is applied. Firstly, the $AR(1)$ model $\tilde{x}(t) = \phi * \tilde{x}(t-1)$ is fitted to the time series $x(t)$, and the residuals $\epsilon(t) = x(t) - \tilde{x}(t)$ are calculated. Then the white noise test is performed on the residuals. If the residuals are white noise, then the underlying process of the given time series is claimed to be red noise.

Defining the characteristics of noise is not necessarily a straightforward task for real-world data. For example, the focus of this study is on atmospheric time series, which are generally characterized by red noise (e.g., Storch and Zwiers, 1999; Ghil *et al.*, 2002; Chen *et al.*, 2013). However, red noise, or an $AR(1)$ process, is frequently fitted to climate time series in order to reproduce the signal, rather than to represent the background noise. It should be recalled here that $AR(1)$ is a stationary linear stochastic process that does not support oscillations (e.g., Storch and Zwiers, 1999). Defining events in the present method as non-$AR(1)$ processes means that the events are non-stationary and/or oscillatory and/or non-linear motions. Other signals end up classified as noise, which can contain physical $AR(1)$-like motions, measurement errors or any other white or red noise signal, none of which are of interest in this context. Additional discussion of these matters is given in section 5.5.

5.2.2 The First Step: Event Detection

The first step locates and detects events by performing a noise test on sliding subequences extracted from the time series. A subsequence is considered to be an event if its characteristics are significantly different from noise. This step is both the major strength and weakness of the method, as it ensures that events can be distinguished from noise even with high noise levels, but it can only be applied if noise separates individual events or at least trains of events in the time series. The procedure is as follows. Using a sliding window, a noise test is performed on each subsequence. A **subsequence** $x_q(t)$ for a time series $x(t)$ with length m is defined as

$$x_q(t) = (x(t_q), \cdots, x(t_{q+l-1}))$$

for $1 \leq q \leq m - l + 1$, where l is the sliding window size, which is also the length of the subsequence. The sliding window sizes l are pre-chosen according to the

scales of interest. For the analysis of multiple scales, various tests have shown that better results are obtained by keeping l constant and block averaging the time series to a desired scale. This is a consequence of the dependence of the test statistic Q on the window length l, and keeping l constant returns consistent results for all scales. After performing a noise test, a test p value is obtained for each subsequence. Assuming the test p value of the qth subsequence $x_q(t)$ is p_q, the result is a p value series: $p_1, p_2, \cdots, p_{m-l+1}$. When the subsequence test p value is smaller than a predefined significance level, we reject the H_0 hypothesis from section 5.2.1. This means that the subsequence's raw data points for white noise test or residuals for red noise test are correlated, which in turn implies that the subsequence is significantly different from noise. Such subsequence is defined as a potential event. If there exists a real event starting at some time point t_0 with the time scale $\triangle_t$, a noise test on sliding subsequences will in general return consecutive potential events from the time point $t_0 - \triangle_{t_1}$ to $t_0 + \triangle_{t_2}$, where $\triangle_{t_{1,2}} \leq \triangle_t$. Therefore, an event is defined only if the consecutive sequence of potential events is long enough. In that case, the central potential event in the progression is chosen to represent the final event, in order to avoid fractional events which do not contain a complete pattern.

More formally, **a potential event** is defined as a subsequence whose noise test p value is smaller than a predefined significant level α. Here we use $\alpha = 0.05$. Assume there exists a consecutive progression of p values $p_s, p_{s+1}, \cdots, p_e$ which satisfies:

1. $p_i \leq \alpha, i = s, s+1, \cdots, e$

2. $e - s \geq l/2$.

Then we define the middle subsequence $x_{\lfloor \frac{e+s}{2} \rfloor}(t)$ as **the event** we are searching for, which is the complete pattern. This definition of the event will be somewhat relaxed when applying the method to complex real-world data.

This step tacitly assumes the existence of noise regions between individual events or trains of events, because otherwise the method could not distinguish between different events. This needs to be considered in applications of the method. The applications to the real-world atmospheric turbulence show that this apparent limitation is not important there, since the time series appear to be composed of intermittent non-AR(1) structures embedded between the regions characterized by AR(1) processes.

In this step, the users need to choose the sliding window sizes l. At present, this choice is subjective and is based on experience and context. In special situations, such as for canopy turbulence, one could use well-established wavelet techniques for determining the relevant time scale (e.g., Collineau and Brunet, 1993b) and choose l accordingly. However, a general recommendation cannot be given at the current level of understanding. As we also see, this step extracts events from time series without organizing them in categories or clusters. This motivates us to design the second step in order to cluster the extracted events for the convenience of comparing and characterizing different types of events.

It should be noted that other techniques could be used for the method's first step. One such example are commonly used wavelet-based approaches for extraction of structures. Wavelets are not used here because they favor large amplitude events or signals, they do not distinguish between signal and noise of comparable amplitude, and they tend to detect structures even when only noise is present in time series (e.g., Collineau and Brunet, 1993b). An example for the latter is that given an artificial time series of linear stationary stochastic Gaussian process without periodicity (i.e., a Gaussian AR(1) process), the wavelet based methods

will find a number of structures, regardless of the fact that the structures are usually considered to be nonlinear, non-Gaussian, etc. While wavelets work well for time series where relatively known structures are present, such as in convective or canopy turbulence, the above mentioned wavelet issues could pose serious limitations in stable situations, or in other applications where building blocks of time series are similarly unknown. Additional discussion about wavelet characteristics is given in section 5.3.2.

5.2.3 The Second Step: Clustering of Detected Events

Clustering is one of the most important tools used by the data analyzers (Williams, 2011). It aims to organize objects into groups such that objects in the same group are similar to one another and different from those in other groups. This is achieved by clustering on the basis of a distance measure between observations. The technique separates data into clusters which are easier for the analyzers to compare and interpret. Hierarchical clustering is one of the most widely used data clustering methods. The idea is to build a binary tree of the data that successively merges similar groups of points according to a dissimilarity measure until all the data are merged into a single cluster. Then the visualization of this tree provides a direct and useful summary of the data. In the end, a choice needs to be made on the number of clusters.

In this step, we use clustering analysis to find the similarity among the events obtained in the first step. In order to account for the global characteristics of the extracted events, a feature-based hierarchical clustering method is used (Wang *et al.*, 2006). In this approach, each extracted event is first described using a feature vector, and then the events are clustered according to the Euclidean distances among the feature vectors, rather than the distances among the raw data of events. The feature set can be chosen for a specific application to best capture

the underlying characteristics of the events. In this paper, the following features of subsequences are considered: standard deviation, non-linearity (Wang *et al.*, 2006), serial correlation (Wang *et al.*, 2006), trend, period, kurtosis, skewness, non-smoothness as well as the maximum, minimum, standard deviation, serial-correlation and kurtosis of the first order difference of the subsequences. The period for the time series $x(t)$ is a revised version of the algorithm in Wang *et al.* (2006) and is determined as follows.

- Calculate the autocorrelation function (acf) for all lags up to 1/3 of the time series length n.
- A local peak is defined at the lag where the acf value is larger than five points before and after it.
- The period is defined as the first peak which is larger than the critical value $1.96/\sqrt{n}$ (Enders, 2003).
- If no peak satisfies the condition above, there is no periodicity in $x(t)$.

The non-smoothness is defined as $\sigma_D/\overline{D}$, where $D(t) = x(t+5) - x(t)$. The other features can be easily obtained using their usual definitions or from the cited references. Besides the above-mentioned statistical features, other features that can characterize events in a specific real world context should also be considered (see section 5.4). To summarize, the second step groups the n_e events extracted in the first step in a d-dimensional (with d being the number of features in the chosen feature set) feature space in order to obtain $k < n_e$ typical clusters of events. In this step, the users need to choose a set of features relevant to their applications, and the number of clusters.

5.2.4 Phase Randomization

At least some of the detected events in atmospheric time series will be coherent structures. Some definitions of coherent structures require the existence of spectral phase correlation (e.g., Provenzale *et al.*, 1992; Gilliam *et al.*, 2000; Chian *et al.*, 2008). As a result, randomization of phase should remove the coherent structures from time series (e.g., Campanharo *et al.*, 2008; Belušić and Mahrt, 2012). Therefore, if the method works properly, it should find considerably more events before than after phase randomization. This fact can be used to validate the first step of the method.

The phase randomization procedure for a subsequence is as follows: (1) Take the Fourier transform of the subsequence to obtain the spectral amplitude and phase. (2) Randomize the phase information by randomly reshuffling phases while keeping the amplitudes unchanged. (3) Use the inverse Fourier transform to return to the time domain. This results in a phase-randomized surrogate of the original subsequence. The results of the method validation using phase randomization are shown in section 5.5.2.

5.3 Application to Artificial Data

Artificial time series are generated with the goal of testing the method in controlled environments. A known number of different structures is inserted in noise of various levels and characteristics. The complexity of noise increases towards red noise, which is a step that leads towards applications to real-world datasets.

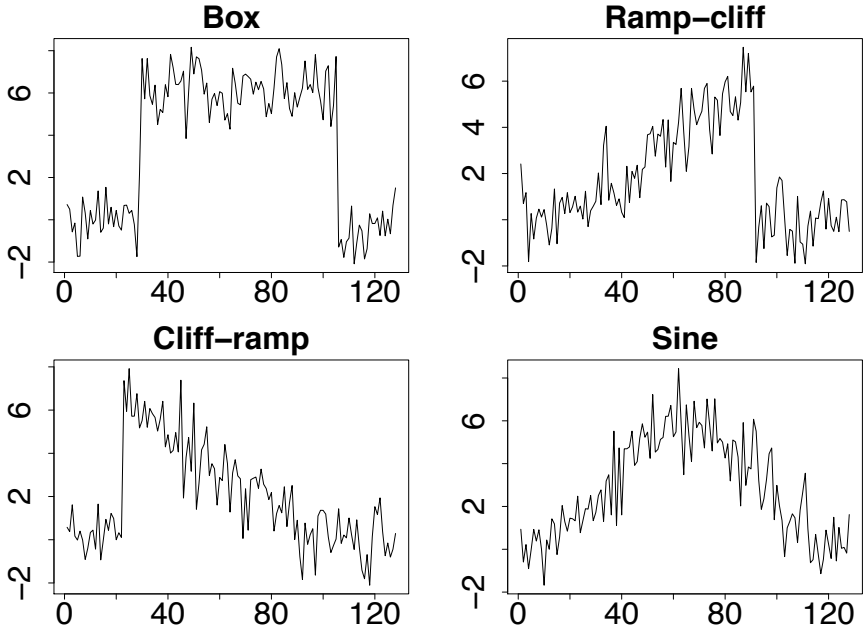

Figure 5.1: *Examples of box, ramp-cliff, cliff-ramp and sine shapes.*

5.3.1 Data Generation

The three basic shapes from the classic Cylinder-Bell-Funnel dataset (Keogh and Kasetty, 2002) are used to create the time series. The cylinder is characterized by a plateau from time a to b, the bell by a gradual increase from a to b followed by a sudden decline, and the funnel by a sudden increase at time a and a gradual decrease until b. Here we call these shapes box, ramp-cliff, and cliff-ramp, respectively, and they represent the typical shapes of structures found in atmospheric time series (e.g., Belušić and Mahrt, 2012), as well as in many other fields. A sine function is additionally included to represent a typical wave signal. The length of the region containing a shape is kept fixed to 128 points. In order to make the task of finding shapes more challenging and hence closer to realistic data, the shapes have variable lengths smaller than 128. The start and end points of shapes vary: a from 16 to 32 and b from 64 to 128. Fig. 5.1 shows an instance of each of the four shapes with some Gaussian noise added.

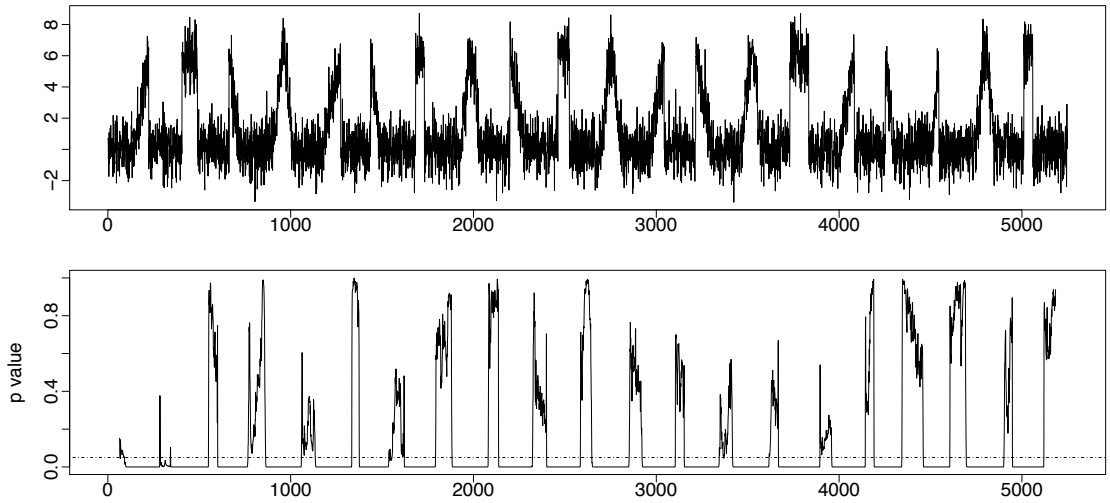

Figure 5.2: *Artificial time series with background white noise with $\sigma = 1$ (top panel) and the corresponding Ljung-Box test p values (bottom panel). The dot-dashed lines represent the threshold $\alpha = 0.05$. A p value smaller than the threshold α indicates a possible shape. Notice that a single p value corresponds to a subsequence of length $l = 128$, and the location of p in the time series corresponds to the central point of the subsequence.*

Time Series with White Noise

Using the four basic shapes, a dataset is generated that contains five instances of each pattern with white noise added as the background. The 20 shapes are distributed in random order, and two neighboring shapes are always separated by a white noise time series with the same length (128). The white noise series is generated using a random number generator following a normal distribution $N(0, \sigma^2)$. An instance of generated artificial time series with $\sigma = 1$ is shown in the top panel of Fig. 5.2.

Time Series with Higher White Noise Levels

The robustness of the method to the level of noise is examined by generating a time series with the level of white noise of $3 * \sigma$. The top panel of Fig. 5.5 shows the artificial time series with noise level $N(0, 3^2)$.

Time Series with Red Noise

As a step towards atmospheric turbulence data, artificial time series are generated with red noise. The first half of this artificial time series consists of four basic shapes and background red noise: $x(t) = \phi * x(t-1) + \epsilon(t)$ where $\phi = 0.4$ and $\epsilon(t) \sim N(0,1)$. The second half consists of two different segments of red noise with equal lengths, where $\phi = 0.4, \epsilon(t) \sim N(0,1)$ and $\phi = 0.8, \epsilon(t) \sim N(0,4)$, respectively. The generated time series is shown in Fig. 5.6.

5.3.2 Results

Background White Noise

In this case we know that the length of the embedded shape regions is 128, so we use a sliding window with the same length $l = 128$ for extracting shapes. In real world cases, l is not known a priori and its values are determined according to the scales of interest. For the second step of the method, the following features are used for this dataset to summarize the extracted shapes: standard deviation, non-linearity, serial-correlation, trend, and maximum, minimum, standard deviation and serial-correlation of the first order difference of the subsequences. The bottom panel of Fig. 5.2 depicts the p value series of the sliding subsequences of length 128 extracted from the artificially generated time series in the top panel. Notice that each shape is related to a sequence of $p < \alpha = 0.05$, so the choice needs to be made about the exact location of the shape. Here we use the middle subsequence according to the definition above. It should be mentioned that although the shapes of the structures are known a priori in this example, the method does not assume that. The latter is important for applications to general real-world cases. The method finds 20 shapes, which are shown in Fig. 5.3. As can be seen from the

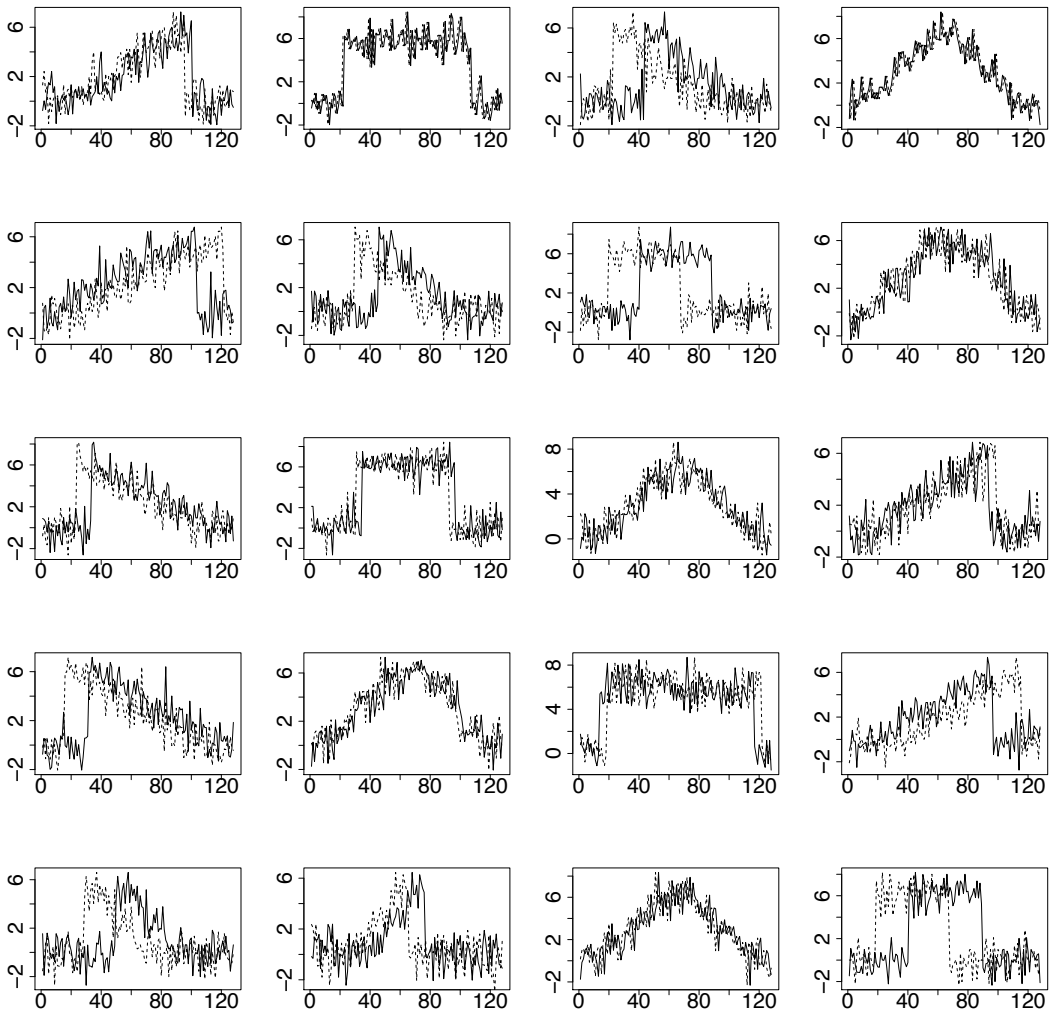

Figure 5.3: *The* 20 *shapes extracted from the artificial time series shown in the top panel of Fig. 5.2. The dashed lines in the background show the original shapes used to generate the time series.*

figure, these are exactly the 20 shapes that were used to generate the time series: five instances of box, ramp-cliff, cliff-ramp and sine shapes.

Once the shapes are extracted, hierarchical clustering is performed on them in the feature space in order to group similar types of shapes together. The dendrogram for the hierarchical clustering is shown in Fig. 5.4 (Wang *et al.*, 2006). It is cut into four clusters since in this case we know that four types of shapes are included in the time series. As the figure shows, same patterns are clearly grouped together, regardless of the differences in lengths or start and end points of shapes.

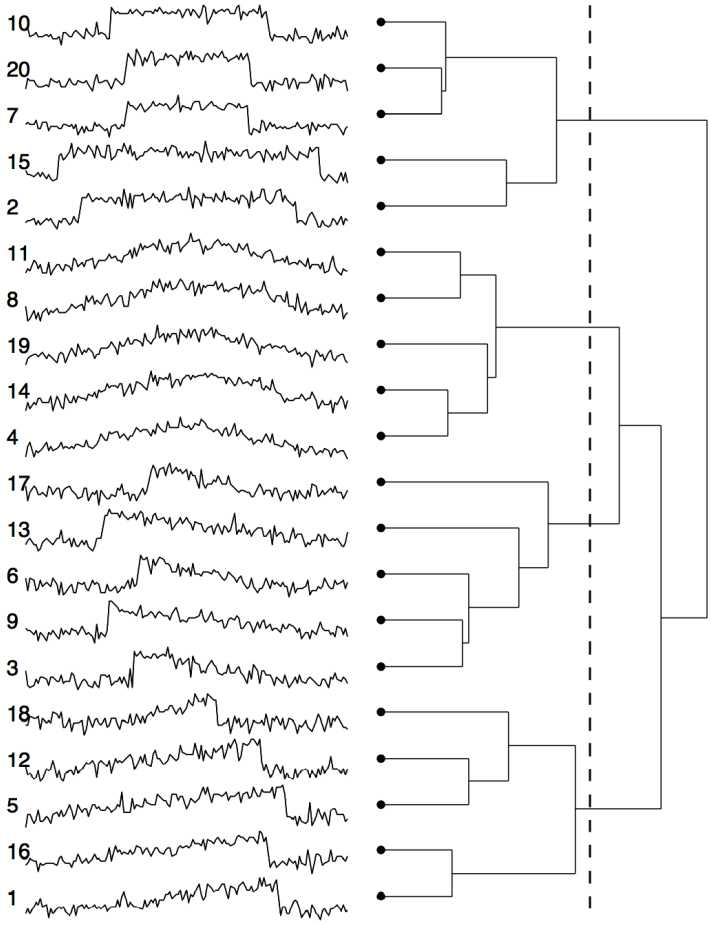

Figure 5.4: *Dendrogram from hierarchical clustering of the extracted shapes based on features; the vertical line shows where the binary tree is cut to get the four basic types of shapes.*

This is one of the highlights of the present approach. It clusters the shapes based on features rather than raw data, which means that shapes with similar characteristics but different lengths or lags are recognized as similar and clustered together, although the euclidean distances based on raw data are large. This is an important advantage of the method when applied to real world data because the shapes in real world are never with exactly the same durations or phases.

Higher Levels of Background White Noise

The above section shows that the new algorithm performs well at finding shapes from artificial time series under a certain noise level. The task becomes more

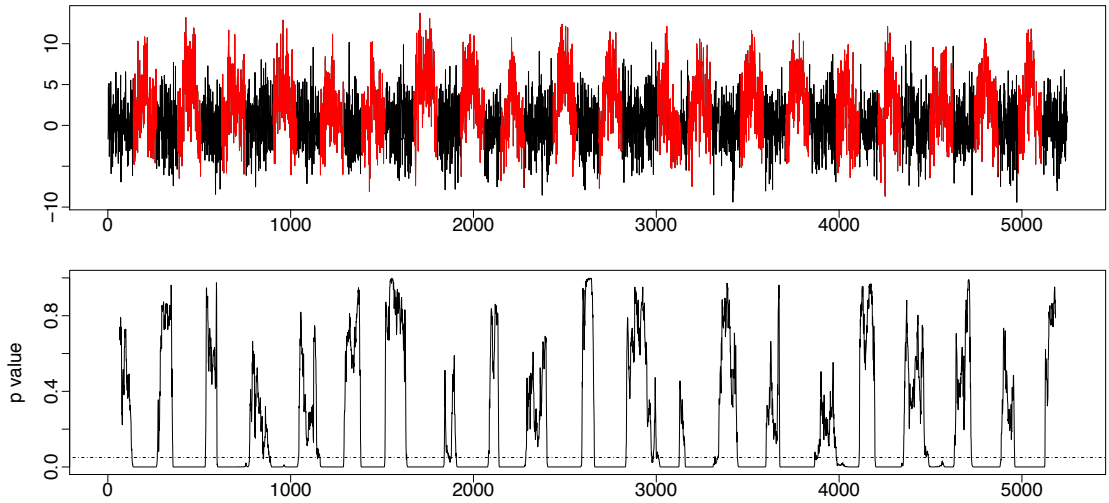

Figure 5.5: *The same as Fig. 5.2, except that the white noise level is increased to* 3 *times as before. The detected shapes are colored red in the top panel.*

challenging with higher noise values because of the difficulties in distinguishing shapes from noise. To illustrate the results, the bottom panel of Fig. 5.5 shows the p value series corresponding to the time series with $3 * \sigma$ noise level. According to the p value series and the definition of shapes, 20 shapes are detected. Even with the magnification of the noise level to three times of the original, the method can still clearly separate shapes form noise. The visual recognition of shapes from the time series would be difficult with this level of noise. The clustering returns the same results as before since the shapes are correctly extracted, so that step is not repeated here.

Background Red Noise

The red noise test is applied to the artificial data with four shapes and the background red noise, shown in Fig. 5.6. The white noise test would not recognize any part of this time series as noise, meaning that the entire time series would be seen as a single large shape. This indicates the importance of correct modeling of background noise before applying the method. Using the present method, the four shapes are correctly detected (Fig. 5.6, top panel). Lower panel of Fig. 5.6 depicts the structures detected by a wavelet-based method that is commonly used for

coherent structure detection (e.g., Thomas and Foken, 2005; Barthlott *et al.*, 2007). The method detects structures at zero-crossings of wavelet coefficients for a certain scale. The wavelet method also finds all four shapes in the first half. However, it detects some noise regions as structures as well. This is particularly evident for the red noise with $\phi = 0.8, \epsilon_t \sim N(0, 4)$ in the last quarter of the time series, which may be confused for structures by appearance. Applying the threshold of 40% of the absolute maximum of the coefficients at that scale, which was introduced in Barthlott *et al.* (2007) for reducing the number of false detections, partially helps by reducing the detection of small-amplitude noise as structures. Regions with larger-amplitude noise are still detected as structures. However, now the third shape, which has a smaller amplitude, is not detected because it falls below the threshold. This example illustrates the benefits of the present method, because it does not depend on amplitudes nor geometries of the signal or noise, but only on the predefined characteristics of undesired noise. The clustering step is not applied here.

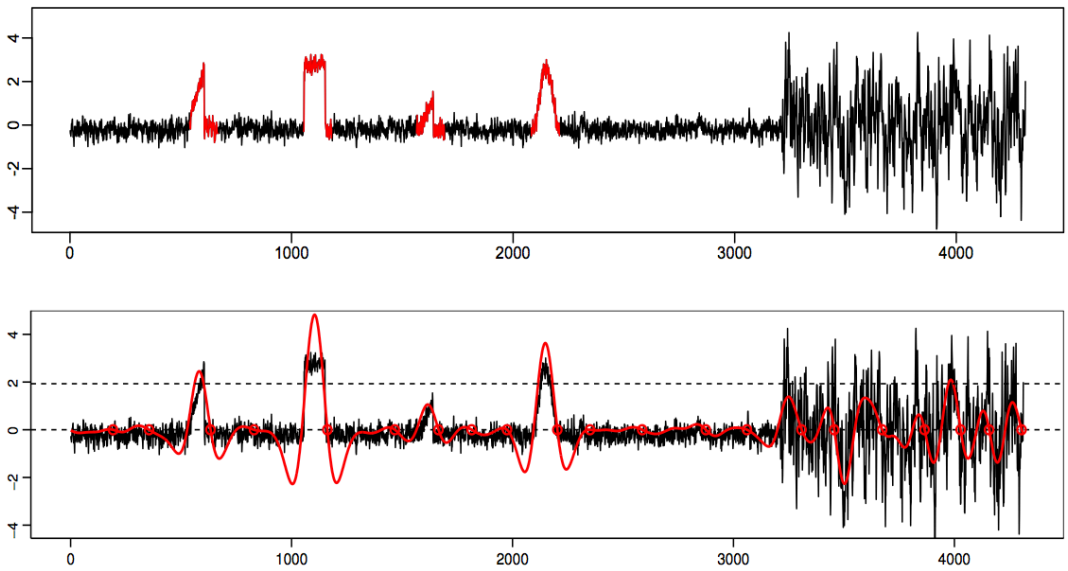

Figure 5.6: *Time series with background red noise and the comparison with wavelets. The first half of this artificial time series consists of four basic shapes and background red noise with the auto-correlation coefficient* $\phi = 0.4$*. The second half consists of two equal-length segments of pure red noise with two different values of* ϕ*:* 0.4 *and* 0.8 *and their* $\epsilon(t)$ *follows* $N(0,1)$ *and* $N(0,4)$ *respectively. The color-coded parts in the top panel show shapes detected using the present method. The bottom panel shows individual coherent structures detected using the wavelets zero-crossing method (open circles) at event duration of* 132 *and the wavelet coefficients (red line). The lower dashed line is the zero line and the upper line indicates* 40% *of the absolute maximum of the coefficients at this scale.*

5.4 Application to Real World Turbulence Data

5.4.1 Data Description

Data from the Cooperative Atmosphere-Surface Exchange Study (CASES-99) are used to test the performance of the method on real-world turbulence. CASES-99 was conducted over a relatively flat-terrain rural grassland site near Leon, Kansas, USA, during October 1999 (Poulos *et al.*, 2002). As a part of the extensive observations, a 60-m tower was equipped with thermocouples at 34 vertical levels (0.23 m, 0.63 m, 2.3 m, and every 1.8 m above 2.3 m) that sampled air temperature five

times per second (Sun *et al.*, 2012), while 20-Hz sonic anemometer measurements were taken at seven levels (1.5, 5, 10, 20, 30, 40, 50, and 55 m).

We use 1-s averages of the thermocouple and sonic anemometer data. The thermocouple at the seventh level (9.5 m) from 11:00 LST of 5 October to 11:00 LST of 6 October is analyzed for extraction, clustering and explanation of shapes of structures. The purpose of using this time period from CASES-99 is to benefit from a number of previous studies that have examined the underlying physical mechanisms of several isolated events on that day. The performance of the method on a real-world dataset is then easily validated by comparing the results with the previous studies.

5.4.2 Event Extraction and Clustering

As discussed before, red noise is used to represent the background noise of real world turbulence data. Accordingly, we use the red noise test for the first step of the method. Faced with the usual case of a consecutive progression of p values $p_s, p_{s+1}, \cdots, p_e$ of the corresponding subsequences, which satisfy the two rules in the definition of the event (see section 5.2.2), the event would be chosen as the middle subsequence for simple artificial data with known lengths of shapes. In the real-world context, choosing only the middle subsequence might result in losing certain parts of the event or train of events. This uncertainty is the consequence of the nonexistence of clear scale separation in the atmospheric flow, whereby events at scales that are somewhat smaller or larger than the prescribed window length l are still significantly different than smaller-scale noise over the range l. So, to take into account a trade-off between not losing events and not keeping too much background noise, in real-world applications we choose the segment from the time point $s + l/4$ to the time point $(e + l - 1) - l/4$, where s is the starting point of the sth potential event and $e + l - 1$ the ending point of the eth. The length of

$l/4$ that is discarded within the first and last potential event was determined by trial and error to avoid keeping too much noise before and after the event. The latter does not impact the final result, because the clustering part of the method is based on global characteristics of events and, as such, it is not influenced by the existence of some noise at the edges of events. With such choice, the window size l is the minimum length of a recognized event, and there is no upper limit to the length of an event.

Using window length l = 120 s (120 points on 1 Hz data), the first step of the method returns 102 events from the temperature time series. Each event is then characterized by a feature vector describing its global characteristics. For this dataset, the following features are used: standard deviation, kurtosis, skewness, period, non-smoothness, and maximum, minimum and kurtosis of the first order difference of the subsequences. Thus the hierarchical clustering algorithm is supposed to cluster the 95 eight-dimensional feature vectors into groups to find similarities among them. However, correlation analysis on the 95 events shows that some of the eight features are correlated, e.g., the correlation between the kurtosis and the kurtosis of the first order difference is 0.91. Therefore, before clustering, we apply the principal component analysis (PCA) to the feature vector to reduce the correlation as well as the dimension. By inspecting the eigenvalues, we choose the first five PCA components to represent the original eight features. Visualization of the clustering is shown in the binary tree in Fig. 5.7. To make the groups clearly separated, the tree is cut into six clusters shown in the six sidebars in Fig. 5.7. The number of clusters was chosen subjectively by visualizing the heatmap and examining the results for several different numbers of clusters.

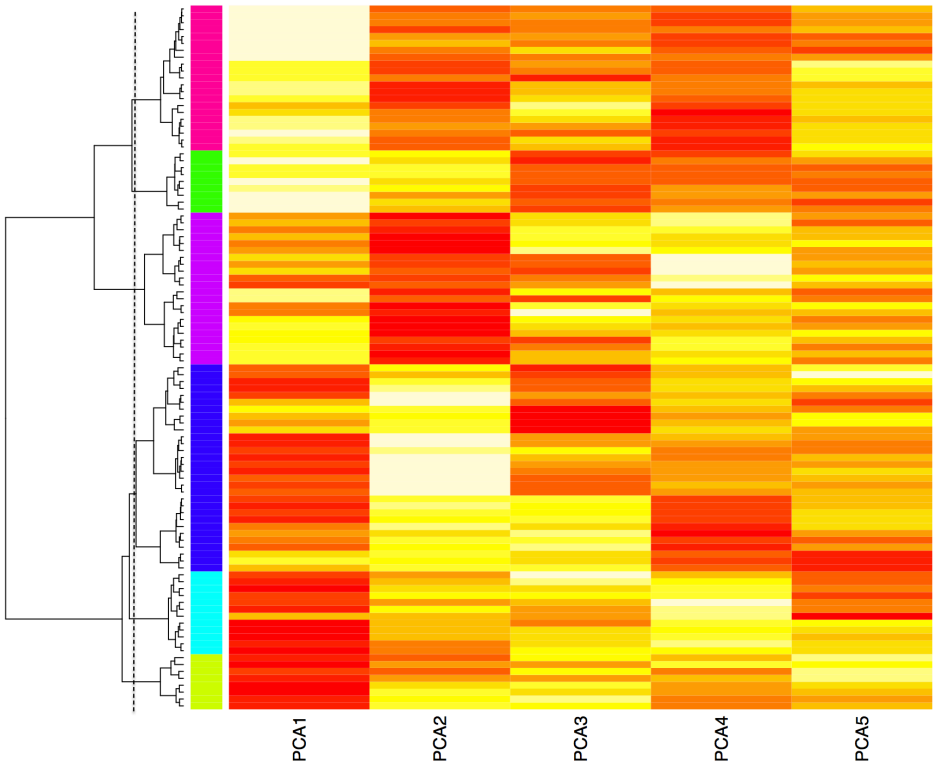

Figure 5.7: *Heatmap for clustering of the extracted events. The hierarchical tree is cut into six clusters represented by the six sidebars. The vertical line shows where the binary tree is cut.*

5.4.3 Characteristics of Events

The following demonstrates the advantages of clustering the events and illustrates that the underlying mechanisms are physically meaningful. Fig. 5.8 shows the transition of cluster numbers for the extracted events, together with the stability associated with each underlying structure. The stability is quantified by the gradient Richardson number $Ri = (g/\theta_0)\partial\overline{\theta}/\partial z(\partial\overline{\mathbf{V}}/\partial z)^{-2}$, where g is the gravity acceleration, θ is the potential temperature, $\mathbf{V}$ is the wind vector, and the overline denotes the time average over the duration of an event. The vertical gradients are calculated using 1.5 and 10 m levels for $\mathbf{V}$, and 0.63 and 9.5 m levels for θ. The time evolution of clusters is related to the evolution of the stability, although the stability is not one of the features used in the clustering procedure. This

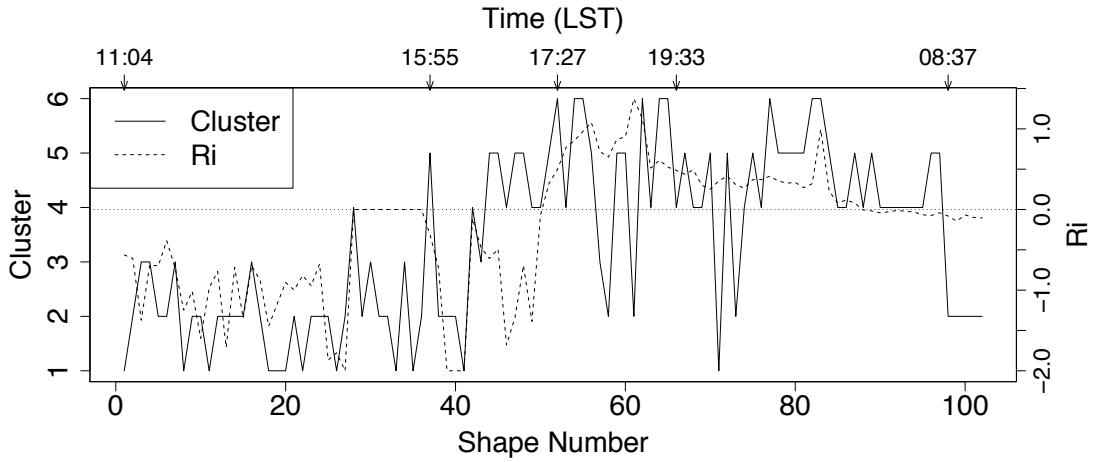

Figure 5.8: *Time evolution of the cluster number and Richardson number of extracted events. The horizontal dotted line denotes Ri = 0. The times on the top correspond to the event times when larger cluster transitions occur. The events were detected in the time series from 1100 LST 5 October to 1100 LST 6 October.*

indicates that clustering is able to group together structures with similar physical characteristics, given only a single-variable time series as the input.

Fig. 5.9 depicts the average depth of structures for each cluster. Since some structures are tilted vertically, the depths are determined by calculating the lagged vertical correlation. The correlation is calculated between the studied thermocouple at 9.5 m and the remaining 27 levels aloft for each event. In order to avoid spurious correlations, the maximum allowed lag, which depends on the event length l_e, is chosen to be $10 * \log(l_e)$. The maximum lagged correlation coefficients are averaged at each height for all events in a cluster, and the average depth of each cluster is obtained as the height where the vertical correlation coefficient falls below $1/e$. Clusters 1, 2 and 6 are characterized by deep events, particularly Cluster 6 where the average structure depth is larger than the tower height so it could not be determined. Combining this information with previous results yields that Clusters 1 and 2 are predominantly composed of deep statically unstable events, while Cluster 6 contains deep stable events. Structures in Cluster 3 are shallow with unstable stratification, and those in Cluster 5 are shallow and

Table 5.1: *Main characteristics of each cluster. The smoothness, defined as* $\overline{D}/\sigma_D$*, where* $D(t) = x(t+5) - x(t)$*, is shown instead of its reciprocal — the non-smoothness defined in section 5.2.3 — for the purpose of legibility.*

Cluster	Ri	Depth (m)	Smoothness	Kurtosis	Skewness	Period (s)
1	-1.07	35.9	3.37	3.85	1.02	31
2	-0.73	35.9	3.78	3.72	0.98	No
3	-0.40	14.4	3.72	9.47	1.82	No
4	0.00	14.4	9.48	3.04	0.29	No
5	0.12	7.2	8.20	3.61	0.62	23
6	0.70	>48.6	12.68	2.26	0.20	No

stably stratified. The distinction between deep and shallow events sustains the usefulness of the present clustering method in that it distinguishes between both the stability and depth of structures even though that information is not fed to the method. It also implies that the characteristics of events in time series carry the information of a wide range of characteristics of underlying structures, which leads to the possibility of classifying and understanding certain atmospheric processes solely from their traces in single-point time series. The latter is clearly true for some specific cases, but is limited for complex three-dimensional motions.

In order to further visualize the clustering results, Fig. 5.10 depicts examples of events in each cluster, and Table 5.1 shows the main characteristics of the six clusters. To summarize, the Cluster 1 examples have the structure typical of periodical deep ramp structures in unstable atmospheric conditions (See Table 5.1). At the same time, Cluster 5 contains periodical but shallow structures in stable conditions. Two of the six clusters, Clusters 2 and 3, contain all the single-cycle ramp shapes in unstable conditions. Fig. 5.9 shows that the ramp structures in Cluster 2 are mostly deep, while those in Cluster 3 are shallow but sharper since this cluster has the largest kurtosis value. Ramp-like shapes in near-neutral

conditions are grouped in Cluster 4. Meanwhile, the most smooth shapes go to Cluster 6. The Cluster 6 structures are deep and apparently wave-like.

A closer look into individual events further demonstrates their physical origin. For example, Fig. 5.11 shows the time-height cross-sections of temperature and vertical velocity for a ramp-like event from Cluster 2. The structure is similar to the one visualized from sodar data using a wavelet transform in Thomas *et al.* (2006). The temperature and vertical velocity are in phase over the tower height, closely resembling the ramps in convectively unstable ABL studied by e.g. Wilczak (1984) and Williams and Hacker (1992). Another example is the event from Cluster 6 that starts at 19:17:06 LST on 5 October. It is a part of the wave-like top-down event studied by Sun *et al.* (2012) that was found to be responsible for turbulence intermittency. The example event for Cluster 5 that starts at 23:39:03 LST on 5 October is again a part of the Kelvin-Helmholtz instablity event that was thoroughly examined in previous studies using other available data during CASES-99, such as radiosondes and a Doppler lidar (e.g., Blumen *et al.*, 2001; Poulos *et al.*, 2002; Sun *et al.*, 2012). Further analysis of physical mechanisms is left for a follow-up study.

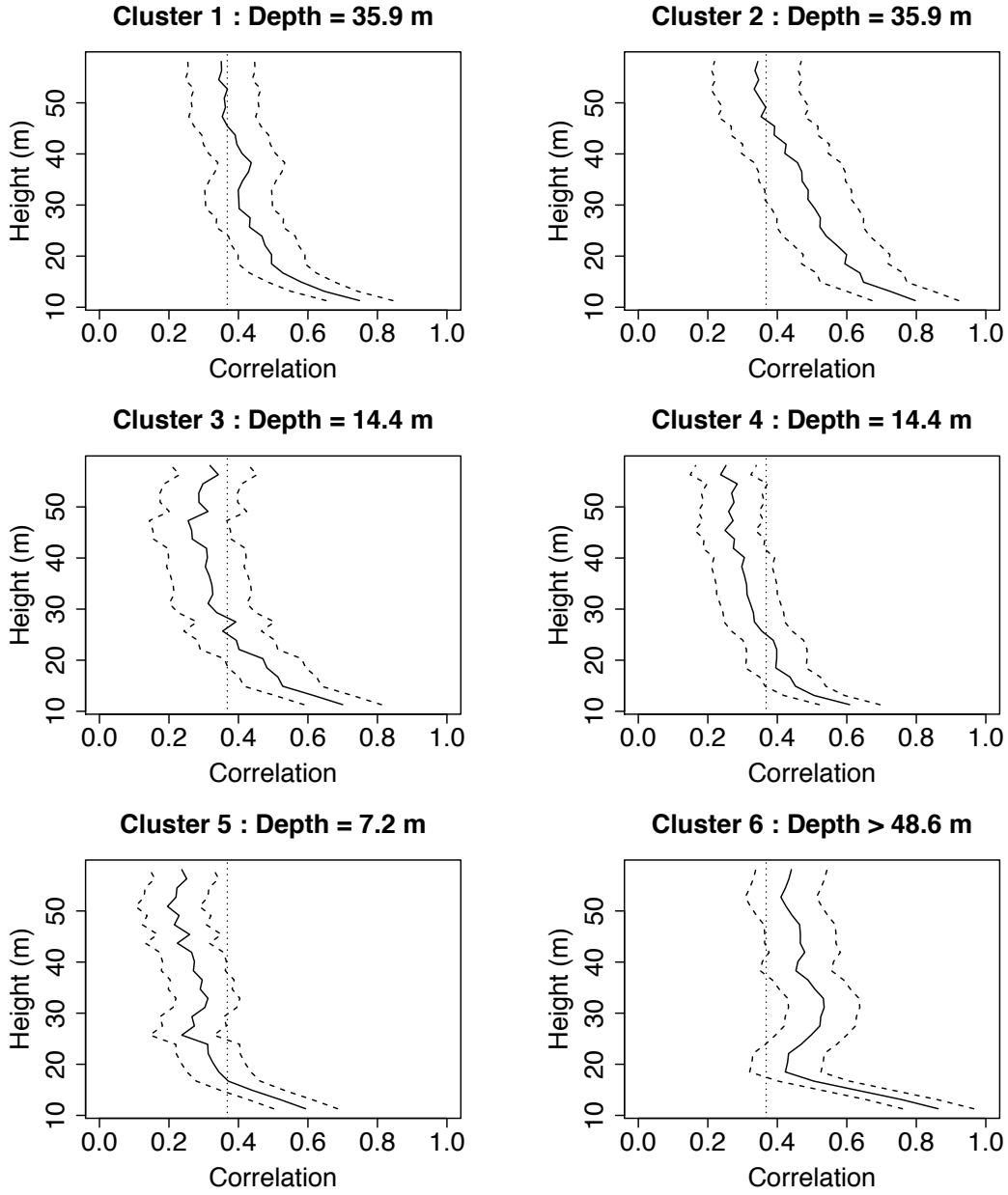

Figure 5.9: *Vertical correlations between the thermocouple at* 9.5 *m and those aloft, averaged over all events for each cluster. The dashes lines show a one standard deviation interval around the mean for each level. The vertical dotted lines represent* e^{-1}*. Titles show mean event depths for each cluster. When the depth is larger than the tower height, it is shown as* > 48.6 *m, i.e. larger than the difference between the highest thermocouple at* 58.1 *m and the reference thermocouple at* 9.5 *m.*

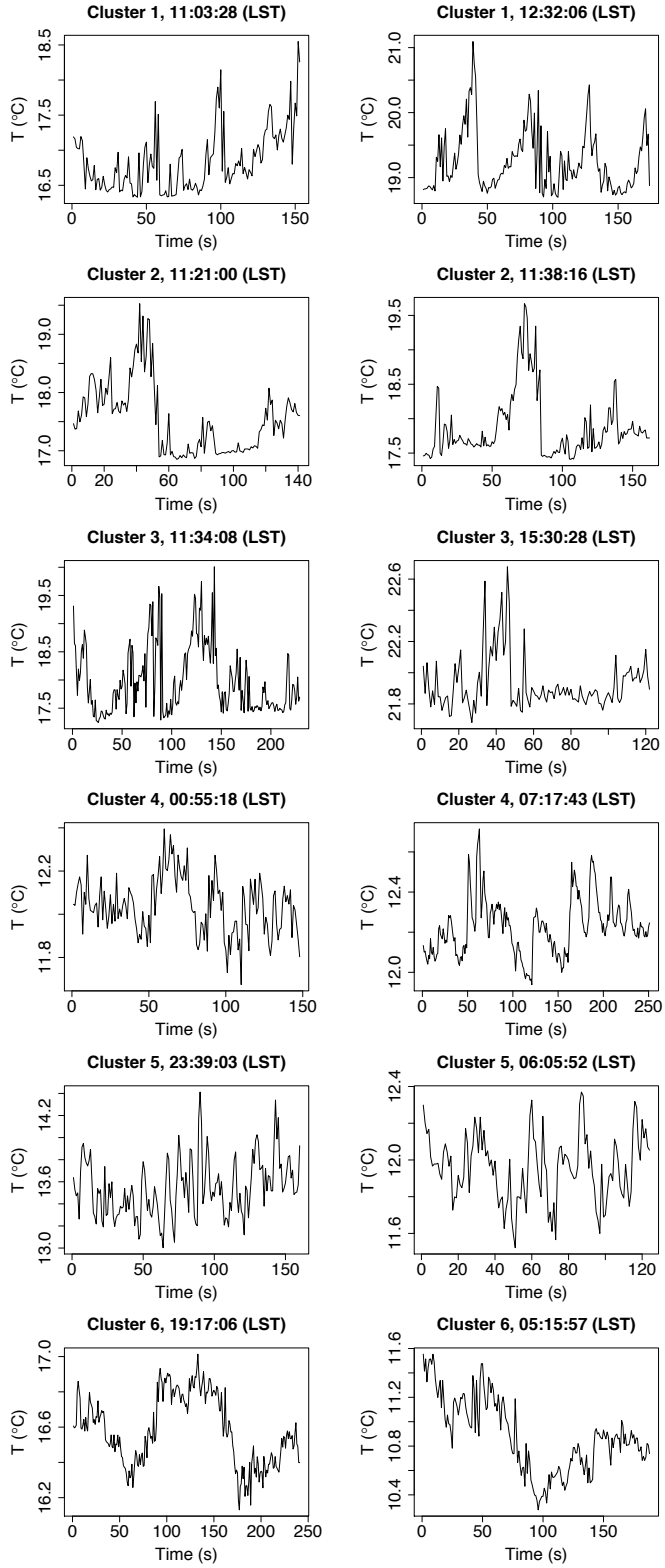

Figure 5.10: *Examples of events from the six clusters: two instances are shown from each cluster. The time of onset of an event is given in each title (the times are between 1100 LST 5 October and 1100 LST 6 October).*

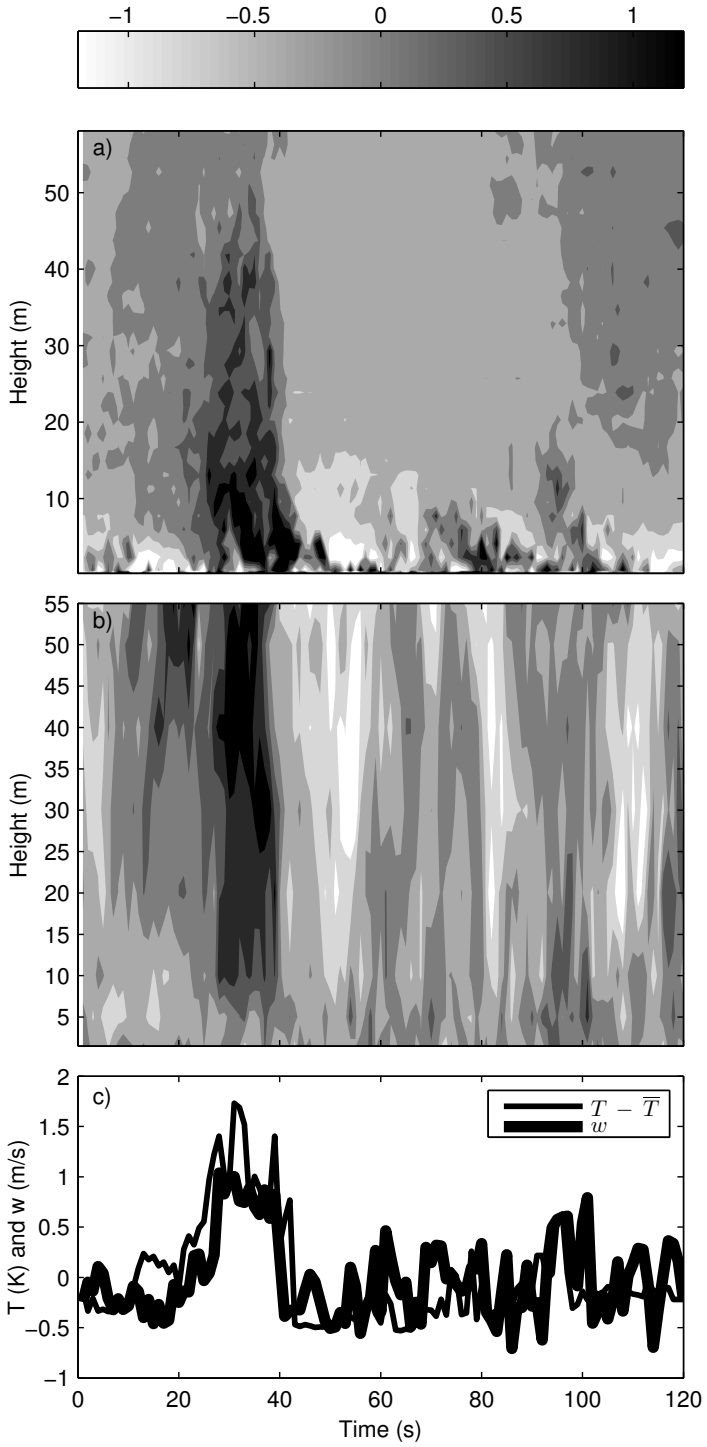

Figure 5.11: *Time-height cross-section of the ramp structure that starts at 11:38:16 LST on* 5 *October showing (a) the temperature perturbation* $(T(z,t)-\overline{T}(z)$*, where the overline denotes the time average over the event duration at each level) from the* 34 *thermocouples and (b) vertical velocity from the seven sonics. (c) The temperature time series with the mean removed of the ramp shape at* 9.5 *m that was recognized by the method. Also shown is the vertical wind speed at the sonic anemometer level* 3 (10 *m).*

5.5 Testing the Event Extraction Approach

An important assumption of the first step of the method in real-world atmospheric application is that an event can be defined as a non-AR(1) process. The suitability of such assumption might not be immediately obvious, so we proceed with two tests that justify this approach. The first test, which is more qualitative, introduces a non-linear component into the linear $AR(1)$ model (Gluhovsky and Agee, 2007) and examines the behavior of the event extraction method. As shown below, the time series generated with higher levels of non-linearity visually exhibit more expressed shapes. The second test investigates changes of event numbers after performing phase randomization (Maiwald *et al.*, 2008) on a real-world time series. The two tests are presented in detail below.

5.5.1 Artificial AR(1) Time Series with a Non-linear Component

We randomly generate 1000 $AR(1)$ time series with the length $l = 500$:

$$x(t) = \phi * x(t-1) + \epsilon(t),$$

where $0 < \phi < 1$ and $\sigma_\epsilon^2 = 1 - \phi^2$, which makes $\sigma_x^2 = 1$. We use $\phi = 0.9$ here to be consistent with the values found from the real world case.

The next step is introducing a non-linear component into the 1000 generated time series (Gluhovsky and Agee, 2007):

$$y(t) = x(t) + a * (x^2(t) - 1),$$

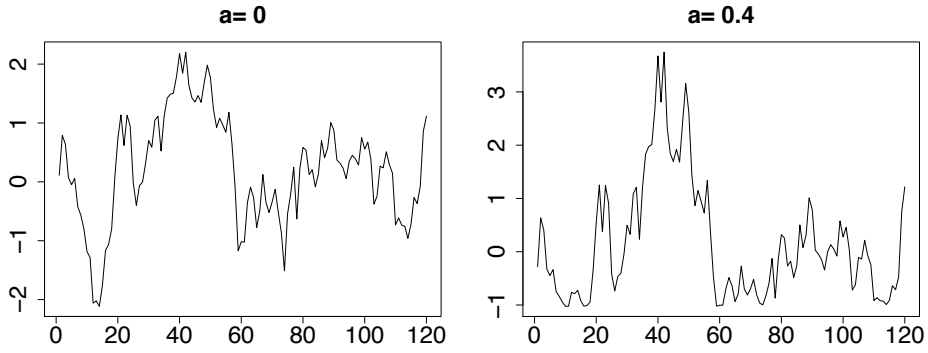

Figure 5.12: *AR(1) time series generated with different values of the non-linearity parameter a.*

where a is a parameter that controls the non-linearity of $y(t)$.

Fig. 5.12 illustrates the changes that occur in the time series as the non-linearity increases. It is clear even from simple visual inspection that individual shapes become more distinguishable with stronger non-linearity. The event extraction method should be able to recognize such differences quantitatively. This is verified by examining the response of the method's red noise test to increasing non-linearity. The percentage of time series with $p > 0.05$ is determined for each value of a, where a ranges from 0 to 0.4 by 0.02. Recall that $p > 0.05$ indicates noise. Fig. 5.13 shows that the percentage decreases with the increase of the non-linearity parameter a. This means that time series become less AR(1)-like as the non-linearity increases, which implies that the method correctly finds more events with stronger non-linearity.

5.5.2 Phase Randomization

As described in section 5.2.4, phase randomization removes coherent structures from time series and can be used to validate the present method. The number of detected events is expected to be significantly smaller in the phase randomized data compared to the original data. It should be noted that the present method

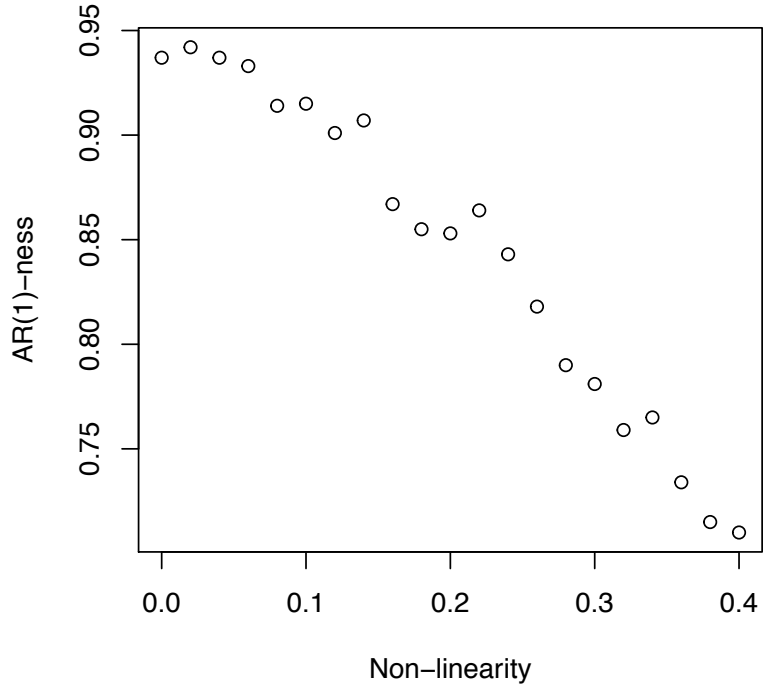

Figure 5.13: *Percentage of the time series characterized by $p > 0.05$ (i.e., that are recognised as AR(1) or red noise) vs. the non-linearity parameter a.*

does not detect only the coherent structures defined in the usual ways. For example, a periodic wave is not strictly a coherent structure because of the absence of phase correlation (e.g., Kuznetsov and Zakharov, 2000), but it is still recognized as an event by our method. In order to alleviate the phase randomization test, we choose a part of the CASES-99 temperature time series during the daytime convective conditions, when the typical ramp-like coherent structures dominate the flow (Wilczak, 1984). The length of the chosen section of the time series is $N = 20000$. The performance of the event extraction method is tested by comparing the number of events obtained from two time series of p values — $p_1(t)$ and $p_2(t)$. $p_1(t)$ is obtained from the unmodified data using the red noise test as in section 5.2.2, while $p_2(t)$ is obtained by phase-randomizing each subsequence before performing the red noise test. The number of events obtained by the method before phase randomization is 26. Using the average over 100 realizations

of phase randomization in order to reduce the uncertainty, only six events are found after phase randomization. This indicates that the method does not falsely recognize events that are not present in the time series.

5.6 Conclusions

A new method for classification of events from time series is developed. The method distinguishes between signal and noise, provided that the nature of the background noise in time series is known in advance. The method is based on two steps:

- A noise test is performed on each sliding subsequence from the time series. The events are extracted as subsequences that are significantly different from noise. This step requires the specification of the characteristics or color of the background noise. Tests are done with white and red noise for artificially generated time series, while red noise is assumed as the model for real-world atmospheric datasets.

- The extracted events are clustered into similar patterns. The second step is based on a set of features that carry the information about global characteristics of an event. This feature-based clustering yields substantially better results than clustering based on raw data.

The method is robust to high levels of noise, which is advantageous regarding the ubiquity of very noisy time series. The application to atmospheric boundary layer time series shows that the method successfully extracts realistic flow structures. The feature-based clustering of the extracted events groups them into clusters with similar physical characteristics, even though the only input into the clustering method is single-variable time series. Finally, the events are detected automatically

without predefining geometries or assuming underlying physical processes. This makes the method a useful tool in exploratory analysis of the dynamics behind time series.

The method is also very flexible and can be tailored to different purposes. The first step can be adjusted to different noise characteristics and the definition of the event can be modified. The second step is highly customizable by choosing different sets of features that are best suited for a specific purpose. The method can be potentially used in areas such as searching for non-noise patterns in solar wind time series (Bolzan *et al.*, 2009), financial time series with underlying red noise (Fu *et al.*, 2001) and other areas concerned with extracting meaningful events from different types of noise.

PART B: Suggested Declaration for Thesis Chapter

Monash University

Declaration for Thesis Chapter 6

Declaration by candidate

In the case of Chapter 6, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
- Developed, established and verified the method - Wrote programming codes and the article	80

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Danijel Belusic	- Provided helpful guidance and editorial work - Contributed through his critical thoughts in atmospheric area	
Kate Smith-Miles	Provided helpful guidance, valuable comments and proofreading	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature	[redacted]	**Date 21/05/2014**
Main Supervisor's Signature	[redacted]	**Date** 21/5/14

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Chapter 6

Classes of Structures in the Stable Atmospheric Boundary Layer

Chapter 6 is based on the article Kang Y, Belušić D, Smith-Miles K. 2014a. Classes of structures in the stable atmospheric boundary layer (Submitted). **Quarterly Journal of the Royal Meteorological Society .**

Abstract. *This paper analyses ubiquitous flow structures that affect the dynamics of stable atmospheric boundary layers. These structures introduce non-stationarity and intermittency to turbulent mixing, thus invalidating the usual scaling laws and numerical model parametrizations, but their characteristics and generating mechanisms are still generally unknown. Detecting these unknown events from time series requires techniques that do not assume particular geometries or amplitudes of the flow structures. We use a recently developed such method with some modifications to study the nighttime structures over a three-month period during the FLOSSII experiment.*

The structures cover about 26% of the dataset, and can be categorized using clustering into only three classes with similar characteristics. The largest class, including about 50% of the events, contains smooth structures, often with wave-like shapes, that occur in stronger winds and weak stability. The second class includes sharper structures with large kurtosis. It is characterized by weaker winds and stronger stability. The smallest class, including about 20% of the events, contains predominantly sharp step-like structures, or microfronts. They occur in weakest winds with strong stability.

Sharper, and particularly shallower, structures are related to transient low-level wind maxima that generate inflection points and may affect generation of turbulence. Furthermore, large wind directional shear, which is another source of transient inflection points, is generated even by deep coherent structures when the background wind is weaker than the structure intensity.

These results show that the complexity of structures can be reduced for the purpose of further analysis using a proper classification. Mapping common characteristics of such

events leads to their better understanding, which, if combined with similar analyses of other boundary layer data, could lead to improving their effects in numerical models.

Keywords. *Submesoscale motions; Coherent structures; Detection of events; Turbulence intermittency; Time series; Clustering*

References are considered at the end of the thesis.

Chapter 6

Classes of Structures in the Stable Atmospheric Boundary Layer

6.1 Introduction

Stable atmospheric boundary layer (ABL) still presents a challenge in all three aspects of atmospheric research: observations, theory and modelling (e.g., Holtslag *et al.*, 2013; Mahrt, 2014). The difficulty of the challenge increases with stability, leaving the majority of the very stable ABL characteristics unknown (e.g., Mahrt, 2014). As a result, numerical models perform poorly in those conditions and the model developers are forced to introduce various artificial enhancements of mixing to keep the models within acceptable performance limits (e.g., Sandu *et al.*, 2013).

A conspicuous characteristic of stable ABL is significant non-stationarity and the associated turbulence intermittency. The turbulent time and length scales can decrease to very small values, leaving a spectrum of generally unknown motions that bridge the gap between the largest turbulence scales and mesoscales. These

motions are usually termed submesoscale or submeso motions (e.g., Acevedo *et al.*, 2013; Belušić and Güttler, 2010; Mahrt, 2009), although some studies additionally use terms such as hybrid motions (e.g., Mahrt, 2014) or dirty waves (e.g., Nappo *et al.*, 2014), to distinguish between different subgroups of motions. The physical mechanisms of these motions are generally not understood, except for a number of individual, usually large-amplitude cases, that were interpreted as Kelvin-Helmholtz instability (Blumen *et al.*, 2001), gravity waves and density currents (e.g., Sun *et al.*, 2002, 2004; Viana *et al.*, 2009, 2010). Otherwise, it is not uncommon to assume that these non-turbulent structures are waves, although distinguishing between waves and other phenomena is not straightforward with the usual limited observations (e.g., Nappo *et al.*, 2014).

Unlike for convective or forest-canopy boundary layers, where many studies have extracted flow structures and analysed their implications for scalar and momentum fluxes, there are only a couple of studies dealing with structures in the stable boundary layer. The majority of these studies examined turbulent coherent structures with time scales on the order of 1 min. Barthlott *et al.* (2007) studied ramp-like turbulent coherent structures over an open field both in unstable and stable surface layer using wavelet analysis. They found that temperature ramp intensity increases with stability and wind shear in stable conditions. Mahrt (2010) extracted large-amplitude step-like temperature structures (microfronts) on time scales of minutes to tens of minutes in stable conditions over a locally flat-terrain site. He classified them into cold, warm and gust microfronts, and concluded that, unlike the typical density currents, cold and warm microfronts are thermally indirect circulations that require an external energy source. However, a myriad of other structures exist in stable boundary layer time series that cannot be detected if only ramp or step geometries are assumed. Their origin, characteristics, and effects are currently unknown, although they could be an important contributor to the overall mixing in stable conditions.

The main goal of this study is to examine whether the stable ABL structures can be extracted and usefully classified into groups with similar characteristics. If achievable, the classification would enable the analysis and understanding of typical dynamical and behavioural patterns in stable ABL for different surfaces and conditions. The measure of usefulness of a classification are groups of similar events that are distinguished from other groups and allow common interpretation of their dynamical and/or structural characteristics. Another relevant criterion is that the classified events should include a significant portion of the total length of the analysed time series. An important constraint is that the stable ABL structures are generally unknown, so the usual methods that are focused on specific characteristics of events, such as their geometrical shapes (microfronts, sinusoidal waves, ramp-cliff patterns, etc.) (e.g., Mahrt, 2010; Belušić and Mahrt, 2012), amplitudes (majority of methods, including wavelets) (Antonia *et al.*, 1979; Wilczak, 1984; Collineau and Brunet, 1993b; Thomas and Foken, 2005; Barthlott *et al.*, 2007), phase relationships (dynamical systems approaches using phase randomization) (e.g., Campanharo *et al.*, 2008; Chian *et al.*, 2008; Kang *et al.*, 2014d), or effects on turbulent fluxes (quadrant analysis) (e.g., Katul *et al.*, 1997), might return biased samples (Kang *et al.*, 2014c). A recently developed method for detecting and classifying events from time series defines events as non-noise time-series subsequences (Kang *et al.*, 2014c). This definition turns the focus towards characterizing the noise, which is typically an easier problem than characterizing events. As a result, the method does not use any assumptions about physical, structural or amplitude characteristics of events, and as such is appropriate for analysing stable ABL time series.

Furthermore, this paper strives to answer the following questions:

- What types of events exist in stable ABL?

- What are the main physical and structural characteristics of different types of events?

- How do these events affect stability and turbulence?

6.2 Data and Methods

6.2.1 Data

The turbulence data were collected during Fluxes over Snow Surfaces II (FLOSSII) experiment conducted in the North Park Basin in north-western Colorado, USA, from 20 November 2002 to 4 April 2003 (Mahrt, 2010). The site is locally flat grass surface frequently covered with snow. The North Park Basin extends about 30 km from west to east and 50 km from south to north, and is located between two mountain ranges extending in the north-south direction with heights of 1000 – 1500 m. Seven sonic anemometers were mounted on a 34-m tower at the following vertical levels: 1, 2, 5, 10, 15, 20, and 30 m. The tower was located at the southern part of a shallow sub-basin that is approximately 4 km wide (Mahrt, 2010).

Quality-controlled 6-s averages of the night-time temperature data from 130 nights at the second tower level (2 m) are used for extraction, clustering and characterization of events. We use 6-s averages for small-scale turbulent fluxes following the discussion in Mahrt *et al.* (2012), where a clear turbulent signal in weak-wind conditions was always present only for time scales smaller than 5 s. Therefore, the priority here is not to include the contribution of all turbulent scales, but to ensure that only the "pure" turbulence contributes to the fluxes in all conditions, including the most stable ones. The fluxes at larger scales are calculated from the 6-s data, with the averaging length equal to the event length. As described in the following subsections, the nominal event length is a constant

parameter in the method (equal to 720 s here), but the final length depends on the characteristics of each event and can become somewhat larger than the nominal event length. As a result, the event-scale fluxes obtained with different averaging lengths include motions on different scales and are not directly comparable. In many cases, there will be considerable contribution of pure turbulent motions to the event-scale flux, particularly for stronger winds, while in other cases the submeso motions will dominate.

Aside from studying the usual variables, we examine the intensity defined as $\Delta\psi_{max} = \psi_{max} - \psi_{min}$, where ψ stands for wind direction (dir), temperature (T), horizontal wind speed (U) or vertical velocity (w). In the case of wind direction, the differences, ranging from 0° to 180°, are calculated between all data points in an event, and Δdir_{max} is taken as the maximum difference.

6.2.2 Event Detection and Classification Method

We use a slightly modified version of the method introduced by Kang *et al.* (2014c), which consists of two steps. The first step detects events from background noise in time series. The second step clusters the detected events, so that similar patterns are grouped together for further analysis of their common characteristics and behaviour.

Step one: event detection

The main task of the first step is separating events from noise. Sequential subsequences of the time series $x(t)$ are obtained using a sliding window with a predefined length scale l. The qth subsequence can be expressed as

$$x_q(t) = (x(t_q), \cdots, x(t_{q+l-1})),$$

where $1 \leq q \leq m - l + 1$, and m is the length of the time series $x(t)$. Events are defined as those subsequences that are significantly different from noise. Note that this step requires specification of the characteristics or colour of background noise. We use red noise, which is typically assumed for atmospheric processes. Red noise is an $AR(1)$ process with positive correlation at unit lag (Chen *et al.*, 2013):

$$x(t) = \phi * x(t-1) + \epsilon(t), \tag{6.2.1}$$

where ϕ is the first-order auto-correlation coefficient ($0 < \phi < 1$) and $\epsilon(t)$ is the white noise process with standard deviation σ_ϵ. Events are detected by performing noise tests on the subsequences. Each subsequence is assigned a p value according to a noise test. When the p value is smaller than a predefined significance level α ($\alpha = 0.05$), it indicates a potential event. The event is detected when a consecutive sequence of potential events is longer than $l/2$ (Kang *et al.*, 2014c). The method follows Kang *et al.* (2014c), with some minor modifications as explained below.

The individual steps of obtaining the p_q value of the subsequence $x_q(t)$ are summarized in Figure 6.1. The original method (Kang *et al.*, 2014c) uses a stationarity test before proceeding with red noise testing. The Phillips-Perron (PP) Unit Root Test (Banerjee *et al.*, 1993; Perron, 1988) is used to test for the null hypothesis that $x_q(t)$ is a unit root process ($\phi = 1$), i.e. that it is non-stationary, against a stationary alternative. This test is used because the AR(1) modelling is only applicable to stationary processes. If the null hypothesis is not rejected, the process is non-stationary. The PP test does not reject the null hypothesis for simple random walk processes ($\phi = 1$), and also for special situations when the process is stationary with a structural break. Simple random-walk processes are not considered as events, but the stationary processes with structural breaks are, so the original method (Kang *et al.*, 2014c) is modified by introducing an additional test to distinguish them — Zivot & Andrews (ZA) unit root test (Zivot and Andrews, 1992). This test allows for a structural break in either the intercept or in the slope of

the linear trend function of the underlying series. Zivot and Andrews (1992) get the asymptotic distribution of ZA test statistic by Monte-Carlo simulation with 5000 replications under the null hypothesis of a random walk. The 1%, 5% and 10% critical values are then obtained from the asymptotic distribution. We draw a conclusion of rejection or non-rejection of the unit root null hypothesis based on the 5% asymptotic critical values. Rejection of the null hypothesis indicates a potential event (stationary process with a structural break). Random walk processes result in non-rejection of the null hypothesis. In our case, when the ZA test statistic value for $x_q(t)$ is more extreme than the 5% critical value, which indicates a potential event, we assign $p_q = 0$, and otherwise $p_q = 1$.

A red noise test is performed on the stationary subsequences obtained from the PP test. The test is based on the definition of red noise given in (6.2.1). We fit an AR(1) model to each subsequence $x_q(t)$ and test whether the model residuals are white noise. If the residuals are white noise, then $x_q(t)$ is considered to be red noise (Kang *et al.*, 2014c); otherwise, $x_q(t)$ is defined as a potential event. The p_q value is the p value of the white noise test for the $AR(1)$ model residuals. Note that since red noise test recognizes a process with a linear trend as a potential event, subsequences are de-trended prior to the red noise test if the goodness-of-fit of the linear model fitted on the subsequence is larger than 0.85. A white noise process can be regarded as a red noise process with $\phi = 0$, so the red noise test recognizes a white noise process as noise as well. Therefore, the rejection of red noise hypothesis means that the process is neither red nor white noise. Further details of the red noise test can be found in Kang *et al.* (2014c).

Finally, whenever there is a consecutive progression of p values $p_s, p_{s+1}, \cdots, p_e$ satisfying

- $p_i \leq \alpha, i = s, s+1, \cdots, e$

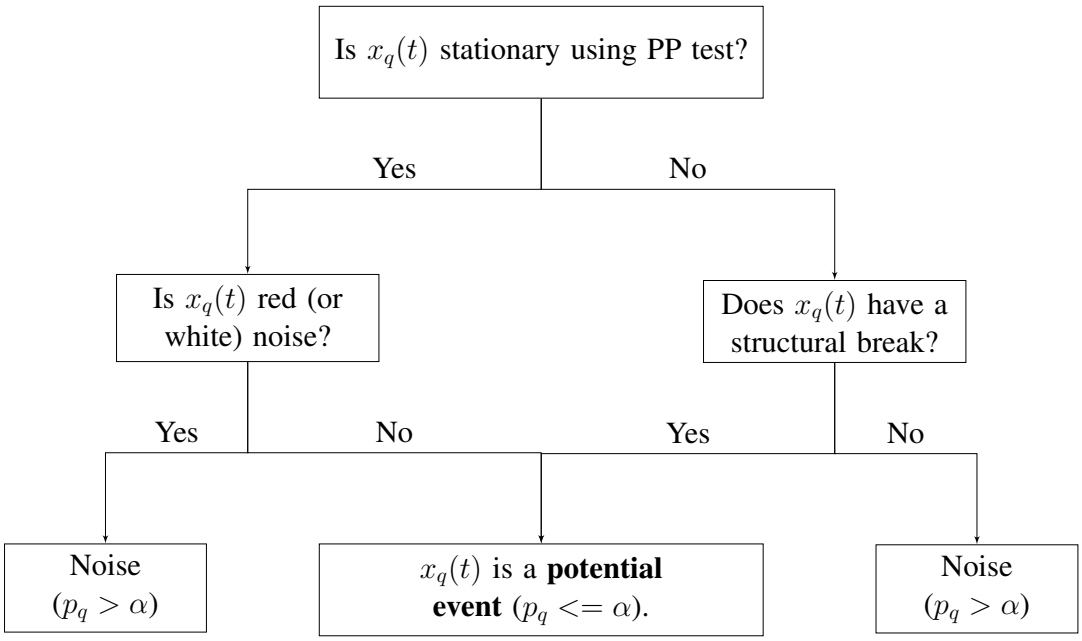

Figure 6.1: *Flow chart of the event detection procedure for a subsequence $x_q(t)$.*

- $e - s \geq l/2$,

the event is defined as the segment from the time point $s + l/4$ to the time point $(e + l - 1) - l/4$, where s is the starting point of the sth potential event and $e + l - 1$ the ending point of the eth. The length of $l/4$ is discarded to remove excess background noise at the start and end of the event (Kang *et al.*, 2014c). Note that using this definition, the event length $l_e \geq l$.

Step two: event clustering

The first step detects the events without organizing them into groups. The second step clusters the detected events to ease the interpretation and understanding of their characteristics. Considering the complexity of turbulence and submeso events, and their variability in length, a characteristic-based k-means clustering method is used in this step (Wang *et al.*, 2006; Kang *et al.*, 2014c). Each event extracted in the first step is firstly summarized using a feature vector. Then the events are clustered in the feature space, in which the cluster prototypes are chosen as the events nearest to the cluster centers. Features used here are standard

deviation σ, kurtosis, skewness, HD (the absolute Difference between averages of the first and second Half), nonsmoothness, test statistic of PP test and ZA test, and maximum, minimum, and kurtosis of the first-order difference of the events. The HD of an event $x_e(t)$ with length l_e is defined as $HD = |H_2 - H_1|$, where $H_2 = \frac{1}{l_e/2} \sum_{i=l_e/2+1}^{l_e} x_e(t_i)$ and $H_1 = \frac{1}{l_e/2} \sum_{i=1}^{l_e/2} x_e(t_i)$. The nonsmoothness of $x_e(t)$ is defined as $\sigma_D/\overline{D}$, where $D(t) = x_e(t+10) - x_e(t)$.

A number of indices have been developed for objective estimation of the number of clusters, each with its own inadequacies. Using multiple indices together and choosing the number of clusters given by the majority of indices (Charrad *et al.*, 2013) has shown as optimal for our purposes.

6.3 Results

6.3.1 Event Extraction and Clustering

Using the window size $l = 120$ (120 points on 6-s averaged data), the first step of the method yields 926 events from the 2-m temperature time series, which accounts for about 26% of the total time series length. In order to establish whether the frequency of occurrence of events changes with atmospheric stability, we calculated the gradient Richardson number (Ri) values for both events and non-events, using $Ri = (g/\theta_0)\partial\overline{\theta}/\partial z(\partial\overline{\mathbf{V}}/\partial z)^{-2}$, where g is the gravity acceleration, θ is the potential temperature, $\mathbf{V}$ is the wind vector, and the overline denotes the time average over the duration of an event or non-event. The vertical gradients are calculated using 1 and 30 m levels. Non-events refer to all remaining parts of the time series after the events are removed. They are divided into segments with the length equal to the sliding window length l before calculating Ri. The events and non-events are grouped together into three stability bins with $0 < \mathrm{Ri} \leq 0.25$,

$0.25 < \mathrm{Ri} \leq 1$, and $\mathrm{Ri} > 1$. The frequency of occurrence of events for a Ri bin is calculated as the percentage of the total time in the Ri bin that is occupied by events. The events account for 19.1%, 42.0%, and 23.5 % of the total time for the three bins, respectively. The events tend to occur with similar frequencies for dynamically unstable ($0 < \mathrm{Ri} \leq 0.25$) and strongly stable ($\mathrm{Ri} > 1$) regimes, but are about twice more common in the intermediate stability range ($0.25 < \mathrm{Ri} \leq 1$). A possible explanation is that for dynamically unstable conditions, the common presence of turbulent mixing transfers the energy towards smaller scales, leaving less organisation and fewer structures at the time scales larger than 10 min that are in focus here. For strongly stable conditions, the lack of instability mechanisms internal to the flow might be responsible for the lower occurrence of events, because then the organised structures have to be predominantly externally generated (such as drainage currents and terrain waves). The high occurrence in the intermediate stability range thus emerges as the favourable combination of the two effects: higher probability of internal instabilities and less turbulence. The twice higher occurrence in these conditions compared to very stable conditions suggests that the structures at the current location are nearly equally produced by internal and external mechanisms. This finding has important ramifications for numerical model performance. While internal mechanisms are to a certain extent taken into account in numerical model parameterizations, the effects of the external unresolved structures are not. Depending on their influence on turbulent mixing and fluxes, the absence of nearly half (all) of the structures could be a significant source of the numerical model under-performance in stable (very stable) conditions.

In the second step, each event is represented using a feature vector. Principal component analysis (PCA) is performed to reduce the correlation among some of the features. The *k*-means clustering algorithm is then performed in the five-dimensional space obtained from the PCA analysis. We obtain the optimal number

of clusters $k = 3$ using the method from Charrad *et al.* (2013). There are 450, 289, and 187 events in clusters 1, 2, and 3, respectively.

Table 6.1 shows the feature values for each cluster. A pairwise t test is performed to validate the distinction among the three clusters for each feature, and almost all the features are significantly different between the three clusters. Such a small number of clusters is somewhat surprising, because it suggests that the complexity of stable ABL over a three-month period could be divided into only three major categories of structures.

6.3.2 Characteristics of Events

Figure 6.2 shows the three nearest and three furthest events from the cluster centroid for each cluster. Combining these examples with Table 6.1, we see that cluster 1 has the smoothest structures, since events in this cluster do not have large sudden change and are stationary (or trend-stationary). Physically, from Figure 6.3 we see that cluster 1 encompasses the events with strongest winds and turbulence. Consequently, they are characterized by the weakest stability (median Ri < 0.25). This agrees well with the analysis of Sun *et al.* (2012), where they indicate that in stable conditions, the wind speed is the predominant factor in generating turbulence, which then acts to reduce the static stability by mixing the vertical temperature differences. This naturally results in small temperature intensities (ΔT_{max}), which implies smooth structures. The wind direction shifts are also small, which could be due to only the inverse dependence of wind direction variability on wind speed in cases when the cross-wind variance does not depend on the mean wind speed (Belušić and Mahrt, 2008), such as for submeso motions (Vickers and Mahrt, 2007). However, the cross-wind intensity (not shown) is also the smallest for cluster 1.

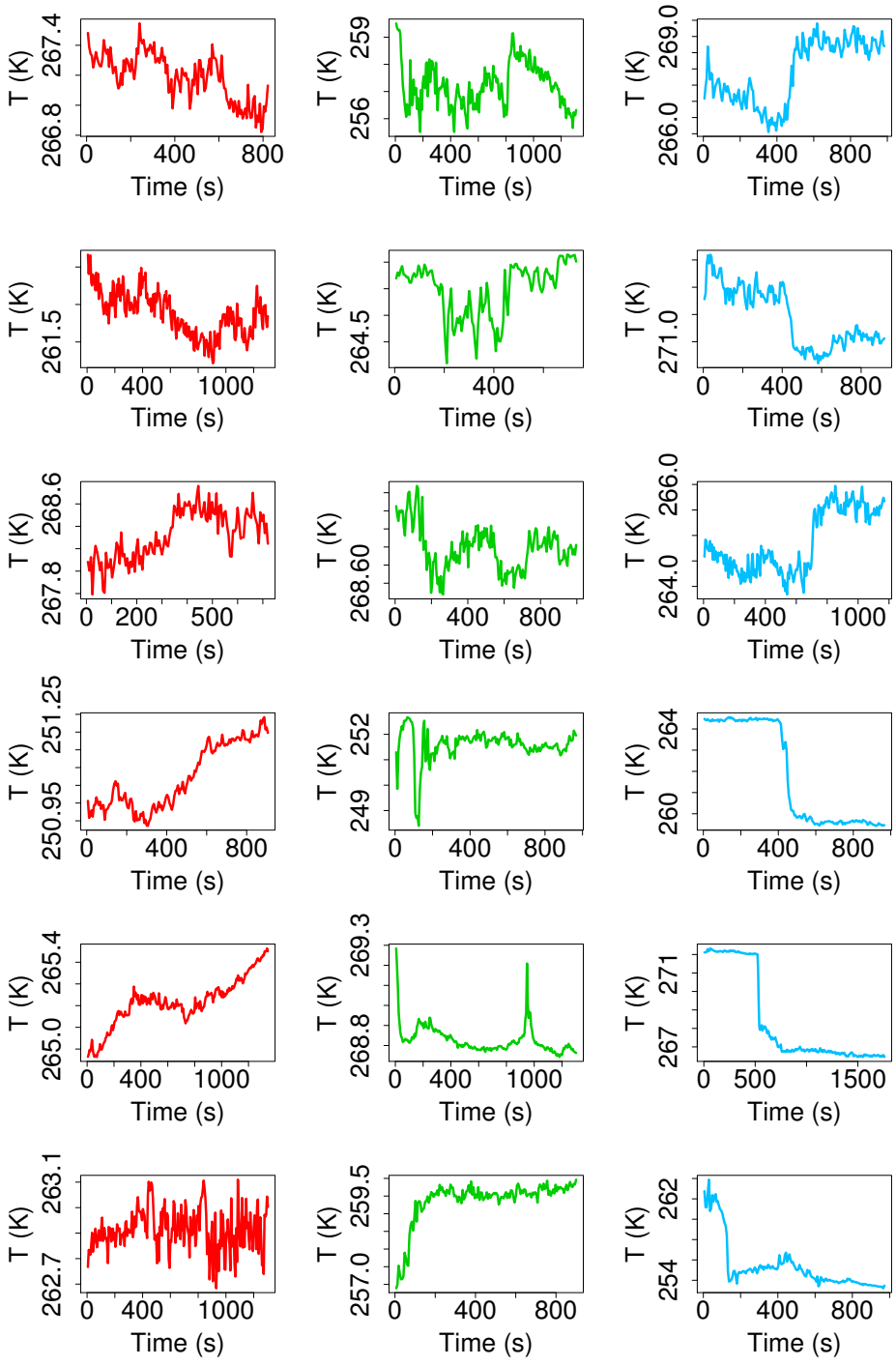

Figure 6.2: *Examples of events in cluster 1 (left panels; red), 2 (middle panels; green) and 3 (right panels; blue). Shown are the three events nearest to the cluster center (top three panels), and the three furthest (bottom three panels).*

Table 6.1: *Centroids of the events in each cluster. The parameters are defined in section 6.2.2. "Diff" stands for the first order difference. The smoothness is shown instead of its reciprocal — the non-smoothness defined in section 6.2.2 — for the purpose of legibility.*

Cluster	1	2	3
Smoothness	0.805	0.738	0.705
σ (K)	0.457	0.435	0.609
HD (K)	0.241	0.133	0.424
Kurtosis	2.459	3.530	1.873
Skewness	0.279	0.639	0.336
Diff max (K)	0.432	0.590	0.426
Diff min (K)	-0.409	-0.548	-0.461
Diff kurtosis	2.959	5.615	6.439
PPstat	-5.378	-3.926	-2.641
ZAstat	-7.191	-5.192	-3.992

Clusters 2 and 3 are characterised with sharp structures that have large skewness and first-order difference kurtosis (Table 6.1). Cluster 3 stands out as being mostly composed of well-defined step-like structures (microfronts; e.g., Belušić and Mahrt, 2012; Mahrt, 2010), while cluster 2 sharp changes are irregular and predominantly short-lived, and may sometimes appear as ramp-cliff patterns. Both clusters have weaker winds, which are associated with weaker turbulent mixing and hence stronger stability, larger wind direction shifts and temperature intensity. The difference is that cluster 3 is farther along the line of decreasing wind speed and increasing stability. The increase in wind direction shifts with cluster number is associated with the increase in the cross-wind intensity. The

common observation is that for non-turbulent scales, the cross-wind variance does not scale with the mean wind speed. We examined the cross-wind variability by calculating the cross-wind variance and intensity for each event using different input data averaging lengths (6 s, 1 min, 5 min), and found consistent behaviour for all scales — the cross-wind variability always increases with cluster number, i.e., with decreasing wind speed (not shown). The examined time scales include the usual submeso scales, so our results differ from Vickers and Mahrt (2007).

The vertical profiles of main physical characteristics of events in each cluster are shown in Figures 6.4 and 6.5. The wind speed on average gradually increases with height for all clusters. Both the wind speed and shear decrease with increasing cluster number. Here the wind shear is defined as the vertical derivative of the wind speed, which allows for negative shear values that point to decrease of wind speed with height. The decrease of wind speed at a certain level indicates the existence of a wind maximum below that level. A low-level wind maximum creates an inflection point in the wind profile within the ABL, which increases the chance for instability and turbulence development. Figure 6.4 shows that cluster 1 has infrequent wind maxima (i.e., negative shear values), which mostly appear at higher tower levels. Wind maxima occur more frequently in clusters 2 and 3. Due to its importance for the development of instabilities, the occurrence of low-level wind maxima is further analysed in the next subsection.

Vertical profiles of the wind direction difference from level 1 are also shown, where the range of differences is [-180°, 180°]. This representation of wind direction simplifies the analysis of the vertical profile, but at a cost of loosing the information about the mean wind direction. The difference also represents a bulk measure of directional shear for each height. Mahrt *et al.* (2013) discuss the influence of directional shear on generation of turbulence in weak-wind conditions. They show that for strong stability, directional shear increases with stability, while the turbulence strength remains roughly constant. This led them to a hypothesis

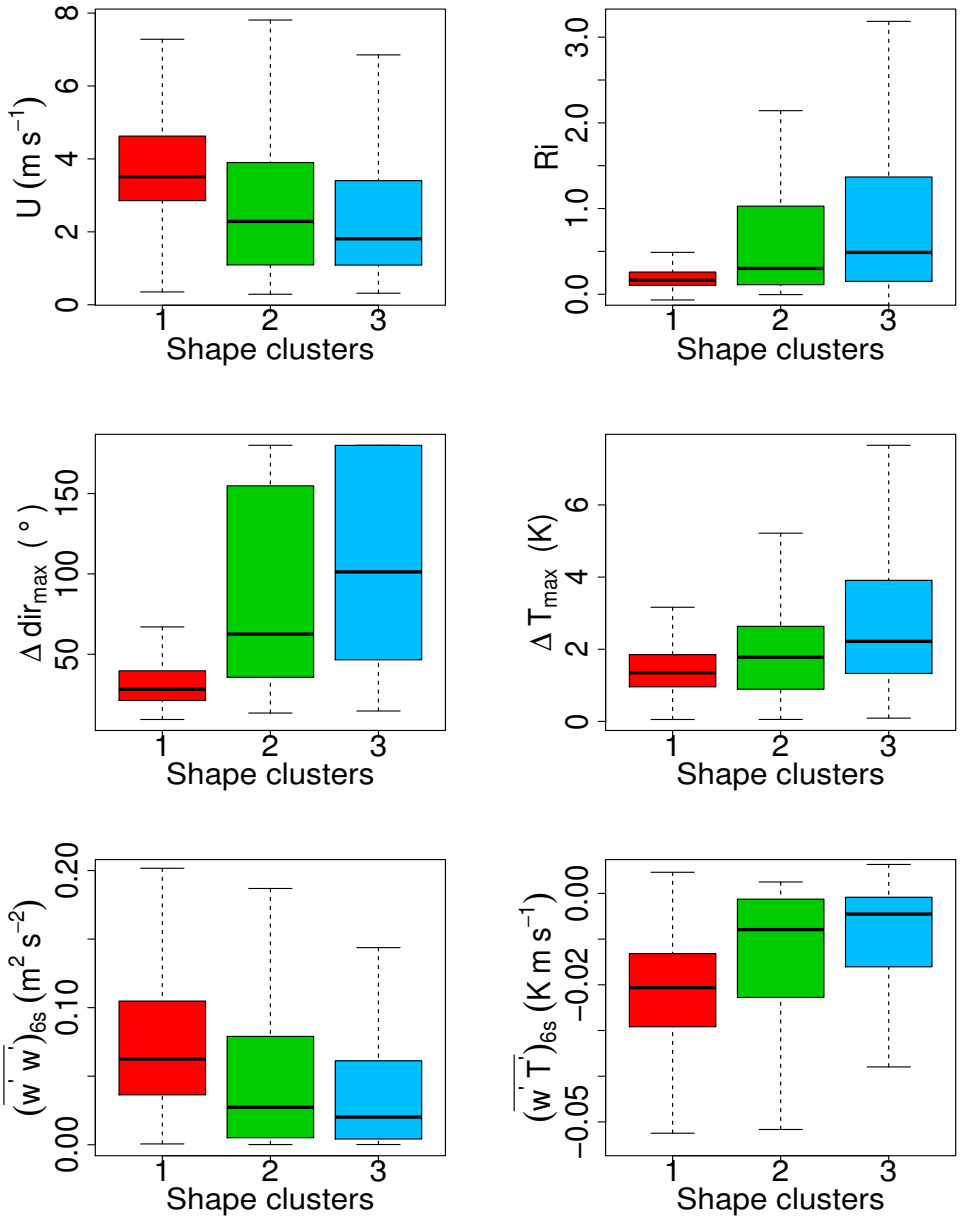

Figure 6.3: *Boxplots of main physical characteristics for cluster 1, 2 and 3 (red, green and blue) at level 2 (2 m). The line in each box represents the median of that cluster, while the bottom and top of the box are the 25th and 75th percentiles. The whiskers extend to the minimum or maximum values within 1.5 times the box height from either side of the box. The subscript '6s' denotes the 6-s averaging interval for the fluxes.*

that the directional shear may offset the turbulence destruction from increasing stability. Figure 6.4 shows that the direction difference is predominantly positive (i.e., clockwise) and increases with height. The increasing clockwise turning with height is consistent with the Ekman model, even though the profiles only reach 30 m above the ground. The direction difference considerably increases with cluster number, both in the median and the spread, which is consistent with the increasing directional shear with stability reported by Mahrt *et al.* (2013).

As already seen from Figure 6.3, the temperature intensity increases with stability. For strong winds in cluster 1, the temperature intensity is almost independent of height, while it decreases with height for clusters 2 and 3. This can be explained if we assume that temperature variability in stable conditions is predominantly generated by vertical displacements. Since the stability decreases with height, the same vertical displacements result in smaller temperature perturbations with height. However, for all three clusters, the vertical velocity intensity increases with height. This increase probably offsets the effects of stability decrease for cluster 1. For clusters 2 and 3, the vertical velocity intensity increase is even larger, but it does not seem to have the same effect on temperature. The latter could be explained if the events are predominantly density currents, in which case the temperature perturbations are caused by impinging air masses that have different temperatures.

The vertical velocity variance increases with height up to 5 m, where it has a pronounced maximum for cluster 1, and approximately plateaus for clusters 2 and 3. This could suggest that the majority of events are associated with elevated turbulence, similar to the upside-down boundary layer discussed in Mahrt (1999). However, Mahrt *et al.* (2013) argue that the ubiquitous increase of $\overline{w'^2}$ with height in stable conditions can be attributed to non-turbulent motions and introduce an alternative measure of "pure" turbulence strength $\overline{w'T'}/\sigma_T$, which they show decreases with height according to expectations. In our cases, this measure has a

weak maximum at 5 and 2 m for clusters 1 and 2, respectively, and for cluster 3 it is constant with height below 2 m and decreases aloft (not shown). This can be explained by the fact that we study mostly non-stationary structures with spatial coherence (including gravity waves, drainage currents, etc.), and as such they will tend to generate shear and instabilities at a certain height above the surface. However, this does not explain the systematic decrease of the downward heat flux with height. For example, in the typical upside-down situations, both the variances and the downward heat flux increase with height or have an elevated maximum (Mahrt, 1999).

The event-scale vertical heat flux has smaller magnitude than the turbulent heat flux, but with much higher tendency towards positive values, i.e., towards countergradient or upward heat flux. This is particularly evident for cluster 3, where the upward heat flux occurs for more than 30% of the events, increasing to about 50% at higher levels. The countergradient heat flux has been associated with various structures previously, predominantly with gravity waves (e.g., Einaudi and Finnigan, 1993; Viana *et al.*, 2009), while our results show that it can occur with the microfronts and other step-like sharp structures. The event-scale vertical velocity variance increases with height, which is probably due to larger vertical length-scales of the perturbations associated with the event time-scales. As a result, these motions become confined close to the ground surface reducing the variance.

6.3.3 Shallow vs. Deep Events

Deep events account for 64%, 57%, and 78% of clusters 1, 2, and 3, respectively. Table 6.2 gives a summary of physical characteristics for deep and shallow events in each cluster. The depths are determined by calculating the lagged vertical correlation of the temperature signal between level 2 (2 m) and higher levels. The

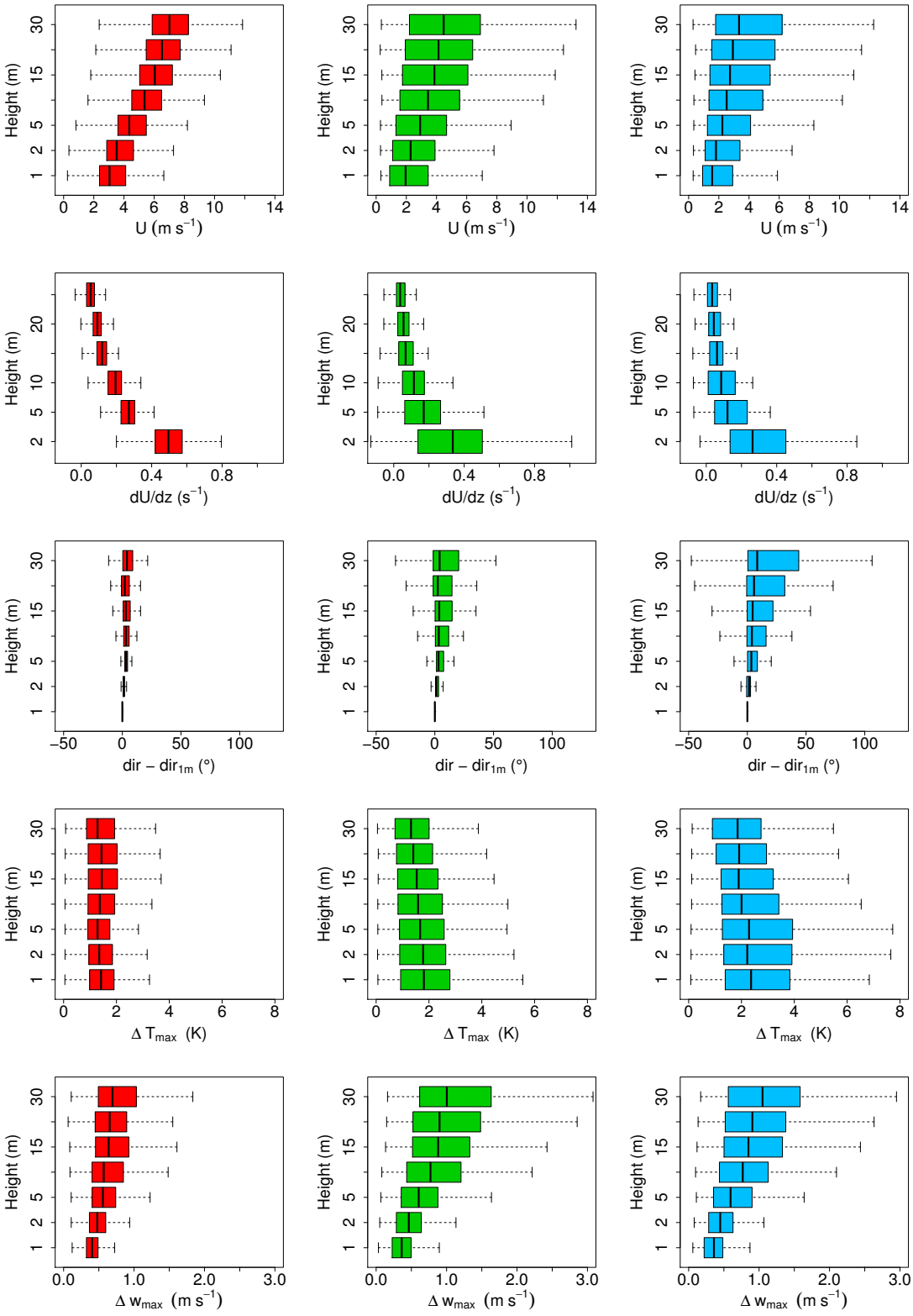

Figure 6.4: *Vertical profiles of the main averaged characteristics of events in cluster 1 (left panels; red), 2 (middle panels; green) and 3 (right panels; blue). dU/dz is the vertical wind shear, calculated as the vertical derivative of the mean wind speed U. dir - dir_{1m} is the mean wind direction difference from level 1 (1 m).*

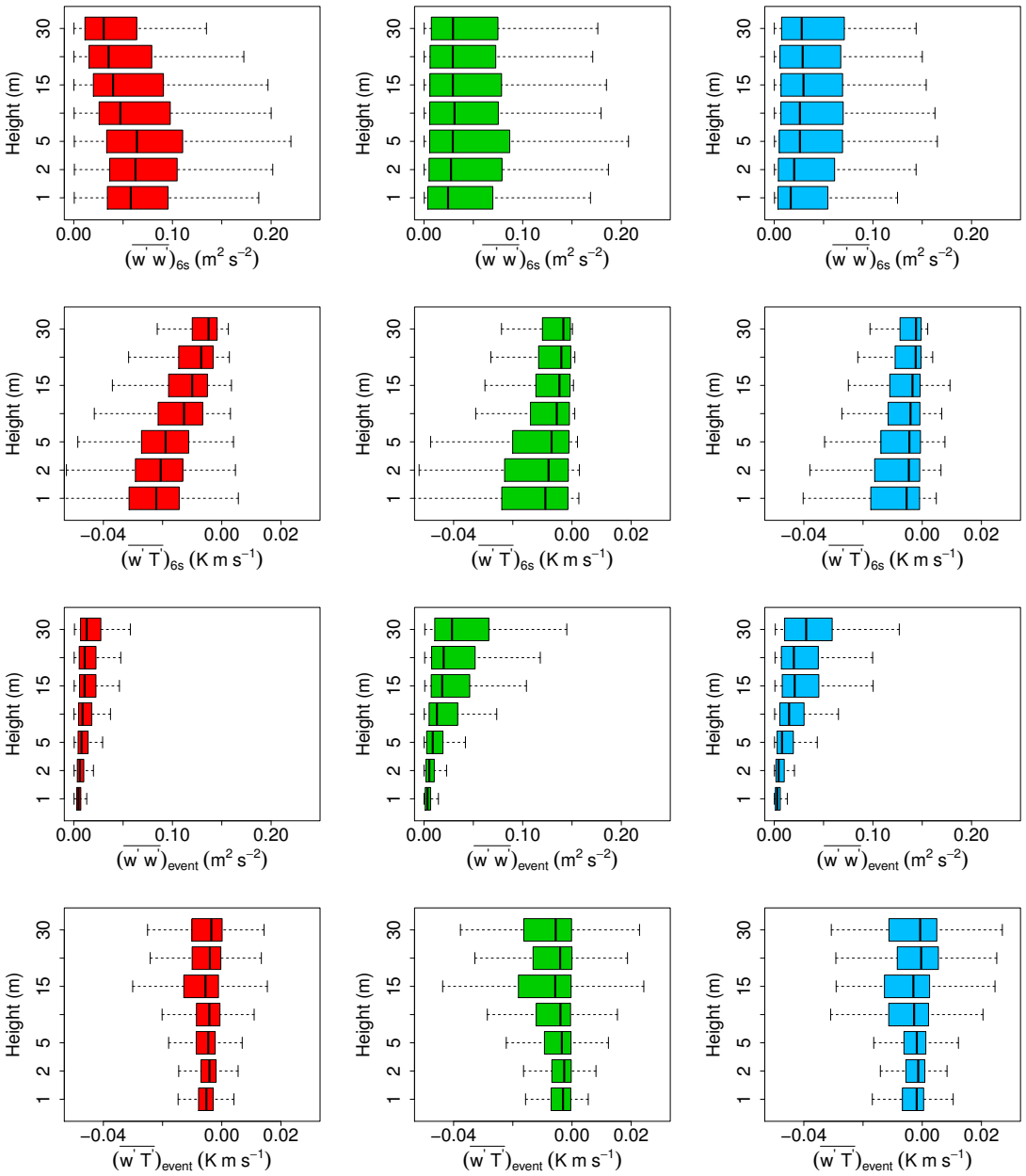

Figure 6.5: *Vertical profiles of the main turbulent characteristics of events in cluster 1 (left panels; red), 2 (middle panels; green) and 3 (right panels; blue). The subscripts '6s' and 'event' denote the averaging intervals for the fluxes (6 s and event length, respectively).*

vertical extent of an event is defined while the maximum correlation coefficient is larger than $1/e$ (Kang *et al.*, 2014c). Deep events are defined as events whose depth is equal to or larger than 28 m, which is the height difference between the top of the tower and level 2, while shallow events have depths less than 28 m.

Table 6.2 indicates that shallow and deep events in cluster 1 have similar characteristics. Since the method clusters the events based on the features such as smoothness, the similarity between deep and shallow events for this strong-wind cluster is consistent with the dependence of smoothness predominantly on wind speed. In other words, large wind speeds occur in cluster 1 regardless of the vertical extent of structures.

Unlike cluster 1, there are substantial differences between shallow and deep structures in clusters 2 and 3, and these differences are similar in both clusters. Shallow structures have considerably weaker winds than deep structures, which is followed by the usual progression of weaker turbulence and stronger stability. However, there are no significant differences between the shallow and deep structures in temperature intensity. This can be explained by smaller amplitudes of vertical velocity perturbations for shallow structures, which coupled with larger vertical temperature gradients for shallow structures results in similar amplitudes of temperature perturbations. Nevertheless, the turbulent vertical wind variance and heat flux are considerably smaller for shallow structures.

The relationship between the event depths and stability for clusters 2 and 3 suggests that the dynamically important structures are predominantly generated and propagate in the ABL surface inversion, rather than appearing from aloft. The ABL height generally decreases with stability as a result of decreasing turbulent mixing. Events that are generated as non-turbulent perturbations of the ABL surface inversion, such as shallow gravity waves, have their vertical extent limited by the ABL height. Density-current types of events, such as microfronts or drainage

currents, depend on the depth of the intruding denser air mass, which is again limited by the ABL height at its origin. The events whose depths most likely do not depend on the ABL surface inversion height or stability are those generated by Kelvin-Helmholtz instability at elevated shear layers. Our results therefore suggest that Kelvin-Helmholtz instability is not the dominant mechanism for very stable weak-wind events. A more extensive analysis is required for proving such conjecture (e.g., Finnigan *et al.*, 1984), but this would require wind and temperature measurements significantly above the current tower height of 30 m.

Another important aspect is the influence of the event depth on wind shear. Figure 6.6 shows percentages of negative dU/dz values for deep and shallow events in each cluster. The percentage of events associated with low-level wind maxima is larger for stronger stability (i.e., for higher cluster number), and for shallower events within a cluster. The percentage also increases with height, as do the intra-cluster differences between shallow and deep events. Note that these low-level wind maxima are not related to the usual persistent nocturnal low-level jet (e.g., Banta *et al.*, 2002), but are associated with the sporadic propagating structures. As the inflection points associated with the low-level wind maxima may lead to more turbulence, the sporadic nature of the structures can result in intermittent turbulence. However, it should be noted that on average, the structures do not increase turbulence. The distribution of effects of structures on turbulence and stability for each cluster is symmetric with near-zero mean — there is about the same number of structures that increase and decrease turbulence or stability. The individual differences in effects of structures are therefore not captured by the current classification. This could be circumvented in future analysis by including other features for clustering, such as the effects of structures on stability and turbulence.

Table 6.2: *Median of main physical characteristics of the deep (shallow) events in each cluster. The subscript '6s' denotes the 6-s averaging interval for the fluxes.*

Cluster	1	2	3
U (m s^{-1})	3.727 (3.176)	2.932 (1.438)	2.032 (1.175)
Ri	0.140 (0.213)	0.193 (0.586)	0.293 (0.907)
Δdir_{max} (^{o})	28.58 (27.85)	52.49(84.71)	84.88 (125.0)
$\overline{w'T'}_{6s}$ (K m s^{-1})	-0.022 (-0.019)	-0.014 (-0.004)	-0.006 (-0.002)
$\overline{w'w'}_{6s}$ (m^2 s^{-2})	0.076 (0.046)	0.049 (0.008)	0.026 (0.006)
ΔT_{max} (K)	1.416 (1.288)	1.720 (1.900)	2.219 (2.215)
Δw_{max} (m s^{-1})	0.519 (0.413)	0.531 (0.366)	0.471 (0.333)
ΔU_{max} (m s^{-1})	2.735 (2.197)	3.021 (1.751)	2.945 (1.799)

6.3.4 Examples of Clustered Events

Typical examples of a deep and shallow event are examined for each cluster. Figure 6.7 depicts two events from cluster 1. The deep event has moderate wind speed, and appears as a wave-like perturbation. Further analysis revealed that the current method recognized only a part of a larger-scale wave with several cycles (not shown). This is a consequence of a given window length for detecting events, and although the final event length is flexible, a structure with a too large time scale might be only partially detected or not detected at all. A remedy for such partial detection is to perform multi-scale analysis using several window lengths. Choosing the largest scale that still recognises an event would encompass the entire event. However, we do not attempt this approach here, because the focus of this study is on a narrow range of scales. The wind speed and temperature wave perturbations are in phase, while the phase of vertical velocity cannot be

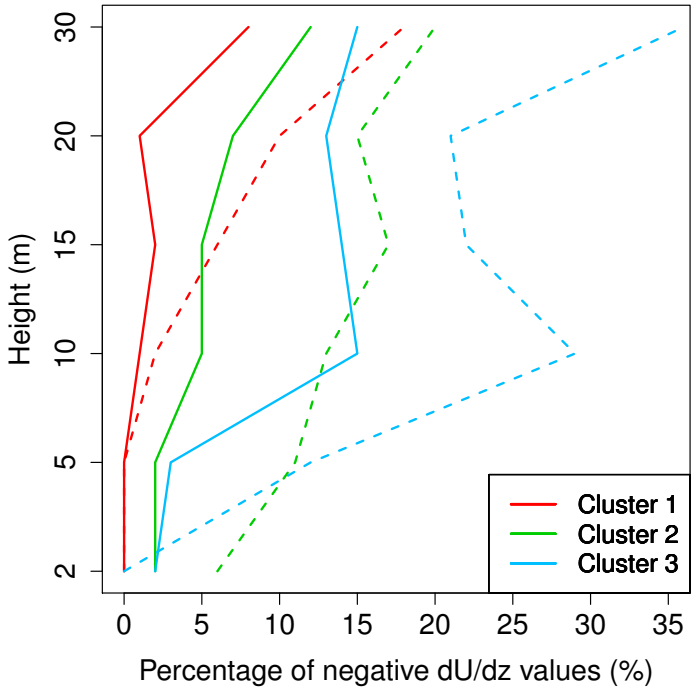

Figure 6.6: *Vertical profiles of the percentage of negative vertical wind shear values (calculated as the vertical derivative of the mean wind speed) for the deep (solid lines) and shallow (dashed lines) events in each cluster.*

determined with confidence. The wind direction is steadily from about 220° (not shown), and the vertical wind direction change is small, on average not surpassing 10° between 30 m and 1 m. The turbulence peaks at 5 m, and is associated with local wave-induced increases of wind speed. Out of the six examples shown here, this event is probably the closest to the typical conceptual consideration of turbulence generated by non-stationary events.

The shallow event in cluster 1 (Fig. 6.7) is a result of an impinging surface-based warm air mass that extends to about 10 m above ground. The wind has a wave-like structure with the maximum speed at 15 m above ground before the arrival of the warm air mass. As in Mahrt (2010), the cold air ahead of the warm air mass rises while the warm air within the air mass sinks, which requires a source of kinetic energy. The arrival of the warm air mass is marked with locally enhanced turbulent

mixing. The maximum vertical wind direction change exceeds 160° between 1 and 30 m during weak-wind periods. Although the temperature event is shallow, the wind structure extends over the tower height without significant change. The structure is superimposed on the background mean wind, which averaged over the event length resembles the Ekman spiral with wind vector turning clockwise with height. The wind turning, together with the low background wind speeds, enable the structure with a relatively small amplitude to diversely affect different levels. At higher levels, the wind direction is steady from about 300°, and the wave-like structure results in quasi-periodic wind minima and maxima. At 1 m, the same structure causes the wind to oscillate out of phase with higher levels, so that the wind speed increases at 1 m while decreasing aloft. This is clearly illustrated by the wind hodographs for different levels (Fig. 6.8a). The hodograph shape that is very similar at all heights veers with height and moves between different quadrants. However, similar shapes in opposite quadrants result in the out-of-phase behaviour, which can generate increased vertical wind speed shear. Furthermore, if similar structures are in neighbouring quadrants, then the wind direction exhibits change with height. If, as in the current case, the oscillation is at the border between two quadrants, then both wind direction and speed may considerably vary with height.

Examples for cluster 2 are two ramp-like structures with the temperature amplitude at 2 m of about 2 K for the deep and 0.3 K for the shallow event (Fig. 6.9). The wind speed and turbulence are also considerably larger for the deep event. The deep event is associated with intermittent occurrences of stronger turbulence, which is generated above the ground and agrees with the concept of upside-down boundary layer. The turbulence increases are associated with the sporadic increases of wind shear that occur without obvious regularity. The vertical wind direction change is small, which seems to be typical for deep events. The ramp-like temperature jump is in phase with the wind speed increase of about 4

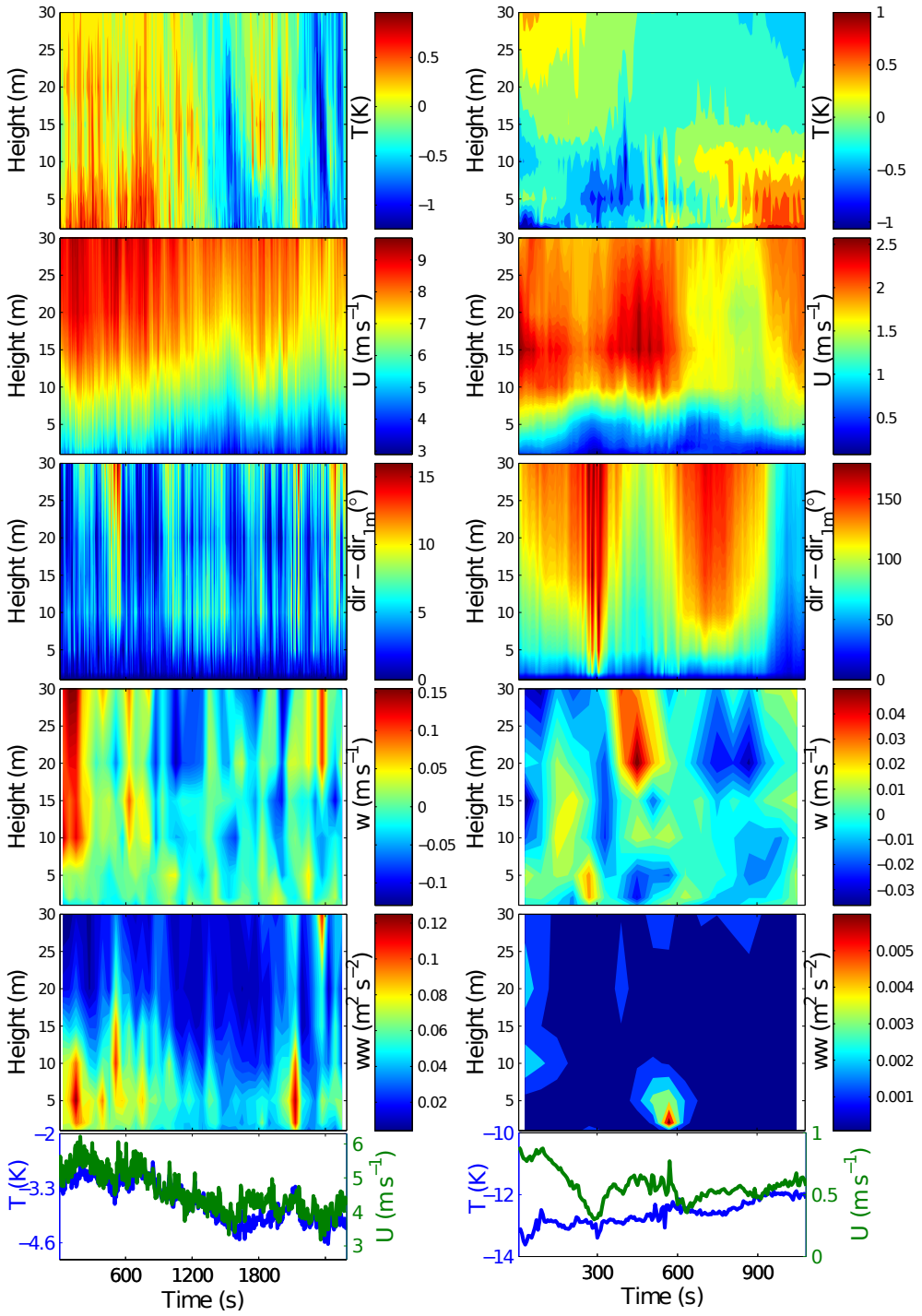

Figure 6.7: *Time-height cross-sections for a deep (left) and shallow (right) event in cluster 1, from the seven levels of measurements. Shown are the temperature with mean removed at each level (T), the horizontal wind speed U, the absolute wind direction difference from level 1 (dir-dir$_{1m}$), the 1-min averaged vertical velocity with mean removed at each level (w), and the 1-min averaged vertical velocity variance at the 6-s time scale (ww). Bottom panels show the time series of temperature T and horizontal wind speed U at level 2 (2 m).*

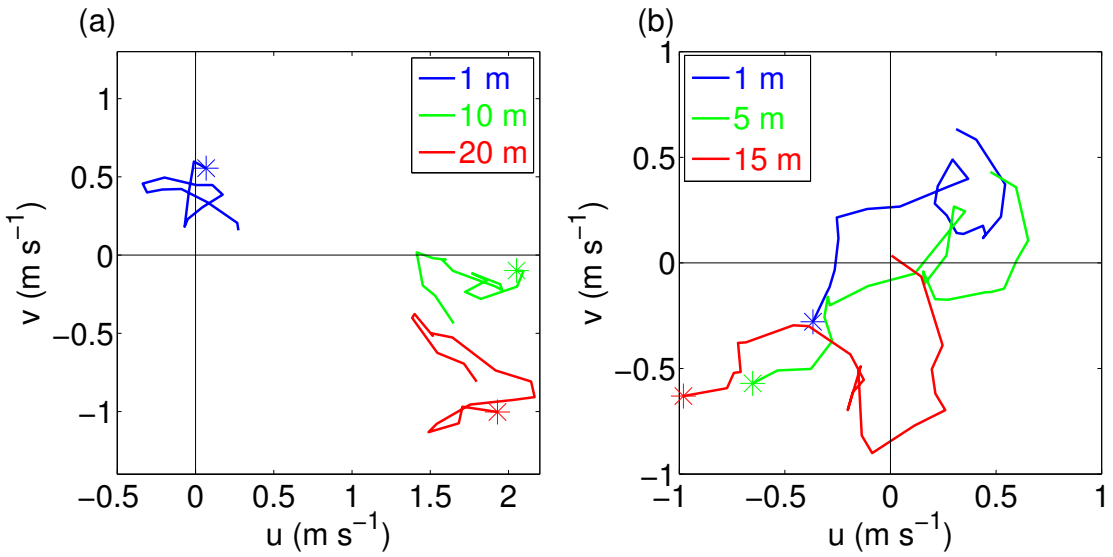

Figure 6.8: *Wind hodograph over the length of the event for (a) the shallow example event in cluster 1 (see Fig. 6.7) at level 1 (1 m), 4 (10 m) and 6 (20 m), and (b) the shallow example event in cluster 2 (see Fig. 6.9) at level 1 (1 m), 3 (5 m) and 5 (15 m). The initial data point at each level is marked with an asterisk.*

m s^{-1}, and is preceded by a deep positive vertical velocity perturbation. The phase relationships suggests a possible wave origin of the event, although this cannot be confirmed since not even a single clear cycle is present. Further analysis would require higher vertical profiles and information about the horizontal propagation of the event.

The shallow event is associated with very weak winds throughout the depth of the tower. The wind maximum averaged over the entire event is at about 15 m, but the time development is irregular and on occasions the maximum is found at or below 2 m. About 500 s into the event, the wind speed has decreased above 5 m, while a short-lived low-level wind maximum appears below 5 m. The wind direction difference between lower and upper levels significantly increases at that time, reaching about 160° between 1 and 10 m. The wind at levels above 10 m is vertically coupled, having approximately same wind direction behaviour, which is consistently different from the low-level wind direction. The hodographs at three different levels exhibit a similar structure, which appears as if resulting from the

same physical process (Fig. 6.8b). However, as discussed above, when the mean wind speed is very low and varies with height, the same or similar hodograph structure may result in significantly different wind direction between different levels. In this case, the same orientation of the hodographs does not even imply the same sense of the wind vector rotation. The wind vector, which has the origin at the centre, rotates clockwise with time at 1 m during the central part of the event, where the vertical wind direction difference is the largest (see Fig. 6.9). At the same time, the rotation at 15 m is counter-clockwise despite the similar structure. The 5-m level is part of a transition zone and behaves as a blend of the levels below and above. The different sense of wind vector rotation at different levels is in this case the cause of the large directional shear. The responsible mechanism is a structure with vertically similar shape that is embedded in the background with low wind speed that changes with height. Note that the initial (i.e., background for current purposes) wind direction is almost constant with height (Fig. 6.9), and the structure has similar shape in the vertical, implying that only the interaction between the two creates the counter-rotating winds at different levels. An interaction resulting with this type of event is much more likely to occur for weak winds. Although the initial background wind direction is almost constant with height, and the structure that affects it is non-stationary, the wind averaged over the event follows Ekman-like turning (Fig. 6.8b).

Both cluster 3 examples are step-like or microfront structures (Fig. 6.10). The first event is a deep warm microfront that is associated with a similar-shaped wind speed structure exhibiting a deep positive jump which is in phase with the temperature increase. This event resembles a class of gust microfronts studied by Mahrt (2010), with the maximum rising motion coinciding with the gust in this case. The structure is associated with a weak clockwise wind turning with time from southerly to more south-westerly direction (not shown), typically resulting from downward turbulent transport of westerly momentum (Mahrt, 2010). The

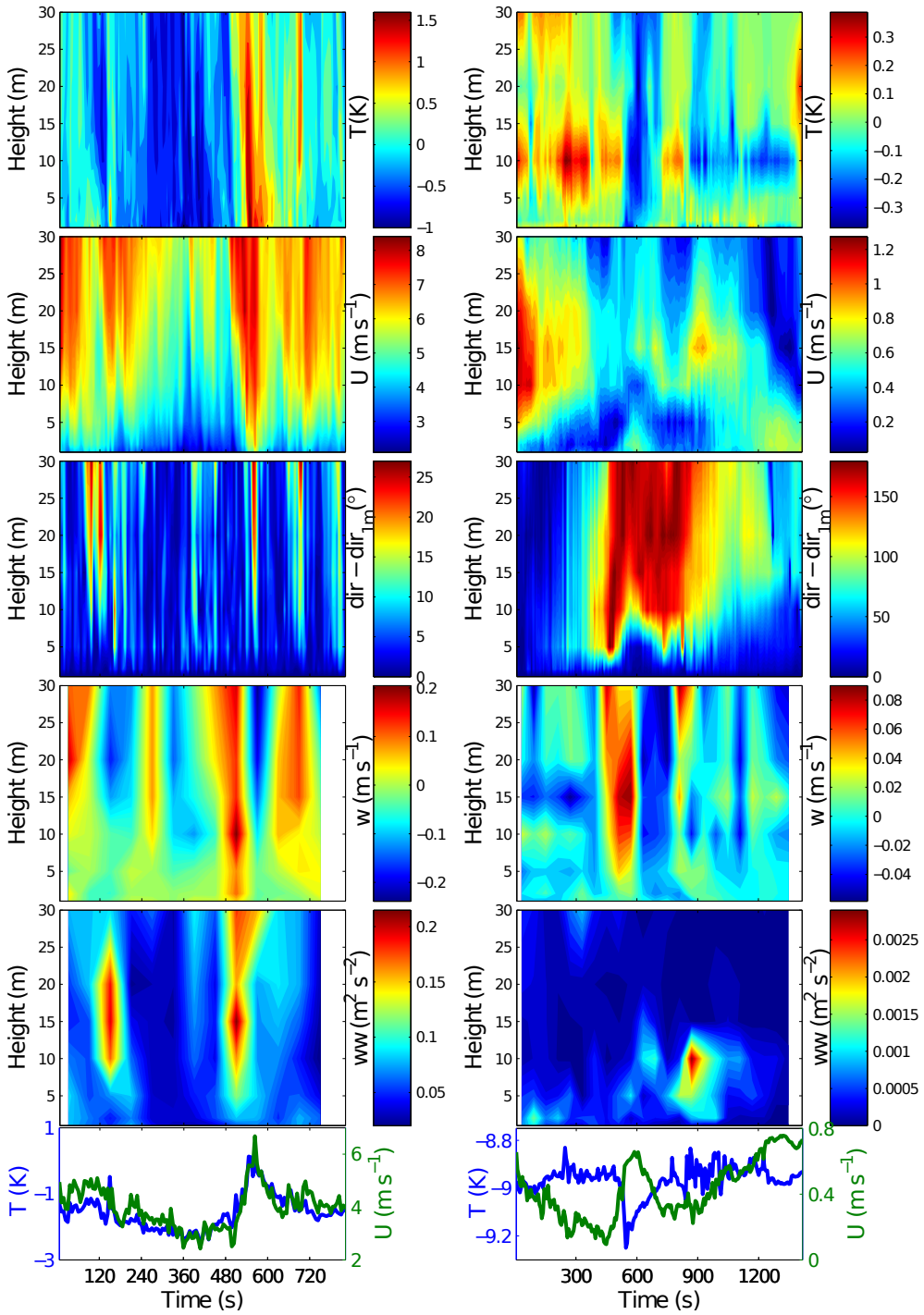

Figure 6.9: *As in Figure 6.7, except that shown are two examples from cluster 2.*

turbulence is mostly generated at higher levels, except after the passage of the microfront where patches of stronger turbulence occur in the first 10 m above the ground.

The shallow example is a typical cold microfront (Fig. 6.10). The surface wind below 5 m suddenly shifts from easterly to northerly flow in phase with the microfront passage (not shown). The source of the cold air is the cold pool located north of the tower (Mahrt, 2010). The wind shift results in a large vertical wind direction change behind the microfront. The wind speed starts decreasing about 2.5 min before the temperature, shortly peaks near the surface at the temperature discontinuity and increases after the microfront passage. The initial decrease of the wind speed does not seem to be related to the surface temperature structure. The strongest turbulent mixing is found at the top of the cold microfront head, resembling other such occurrences (e.g., Hohreiter, 2008).

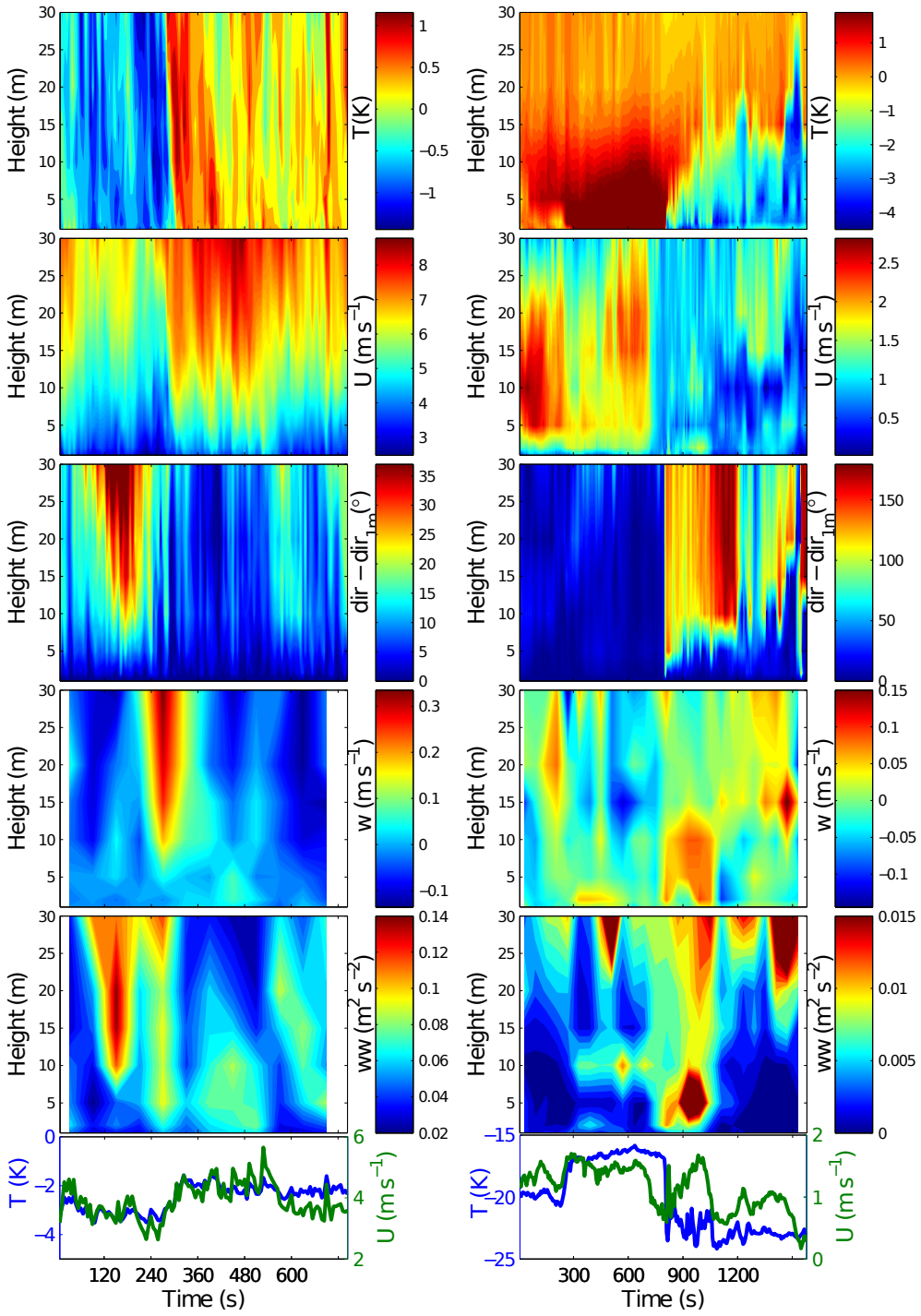

Figure 6.10: *As in Figure 6.7, except that shown are two examples from cluster 3.*

6.4 Conclusions

Events in the stable ABL were detected and classified using a recently developed method (Kang *et al.*, 2014c) with a few modifications. The detected events were categorized into three broad groups or clusters, although they span a wide range of different structural shapes, intensities and background conditions. The current classification using clustering is based only on statistical measures of characteristics of events from time series. Yet, the similarities within and differences between such clusters are associated with corresponding dynamical similarities and differences. The first cluster encompasses smoothest, sometimes wave-like structures that are associated with largest wind speeds, strong turbulence and weak stability. The second and third clusters have predominantly sharp structures, which are step-like (microfronts) for the third cluster. They have weak winds and turbulence, strong stability and are associated with large wind direction shifts.

The vertical structure of clustered events shows that the occurrence of low-level wind maxima increases with stability, and is more likely for shallow events. These wind maxima are mostly non-stationary, introduce inflection points and thus can affect the generation of intermittent turbulence. Large wind directional shear, which can be another source of inflection points, can occur also when deep coherent structures modify the weak background wind profile. The individual examples examined here were associated with intermittent turbulence that can be generated either at higher levels, corresponding to the upside-down boundary layer, or near the surface, although in that case peaking at about 5 m above the ground.

None of these phenomena are resolved or taken into account in numerical models, except through artificially increased mixing. This study provides an avenue for addressing the complexity of the events and their characteristics, which can help

in obtaining their systematic physical or statistical description and consequently improving their treatment in numerical models. More and different data is needed for the latter, which can be achieved by extensive measurements of different stable boundary layers using new techniques and instrumentation.

We note that further improvements can be made to the current method. For example, the multi-scale nature of atmospheric data calls for a range of window lengths for the detection of events, which comes with the unavoidable overlap between events at different scales that needs to be addressed. Furthermore, to cluster events, the method uses generic time series features that are not specifically tailored to the nature of atmospheric time series, and future work could investigate more problem specific features.

Chapter 7
Conclusion

Chapter 7

Conclusion

7.1 Contributions of the Thesis

In this thesis, the aim is to develop a new method for detecting and classifying structures from turbulence time series. In the literature of turbulence event detection, researchers usually assume structures have certain geometrical shapes or other physical properties and detect them by finding the predefined patterns. However, these predefined patterns may not fully explain turbulence phenomena. The emphasis put on certain types of structures leaves others unidentified. This thesis contributes to the literature by proposing new methodologies that can detect and classify flow structures without those assumptions. Through the application of these new methods to artificially generated datasets and real world datasets of interest, it is empirically shown that the new methods can yield realistic results and they appear able to provide substantial insights into the understanding of turbulent transport processes.

In Chapter 2, a real-time change detection method is proposed by which we can visualize how structures transit from one cluster to another. Results on synthetic

and benchmarking datasets are consistent with the real world explanations. However, when we applied the method to turbulence data, it did not work as well as for the synthetic and benchmarking data. The main reason lies in the complexity of turbulence time series. That is also why so little literature could be found about this topic in atmospheric science.

In Chapter 3, it is found that the space and time organized structures in turbulent flow do not necessarily have correlated phases. From both measures used to quantify the phase correlation, a significant proportion of the structures detected using wavelet-based method from the thermocouple temperature time series are weakly phase-correlated. A number of examples of space and time coherent structures with weak phase correlation are presented. The results warn about the vague terminology and assumptions around coherent structures, particularly for complex real-world turbulence. Furthermore, this study indicates there is great uncertainty in the definition of structures.

In Chapter 4, the main contribution is proposing a two-step method for event detection and classification, which shifts the focus from defining structures towards defining noise (i.e., non-events). In the first step, it ignores the noise part in a time series by performing noise tests and focuses on the non-noise subsequences, which improves the meaningfulness of the event searching procedure by avoiding the meaningless limitation of subsequence clustering described in Keogh *et al.* (2003). In the second step, performing clustering in the feature space keeps the information of the main characteristics of event shapes and has been shown to yield better results than the clustering based on raw data. Furthermore, it improves the interpretability of clusters. Experimental results on synthetic datasets show that events used to generate the data can be exactly detected and clustered. It has been shown that the proposed method is robust to higher noise levels, which is a strong advantage regarding very noisy time series like turbulence. More importantly, events can be detected without predefining geometries or assuming underlying

physical processes. Results on the real world time series show that the method has great potential for application. We left further research and analysis on real world time series to the following chapters.

In Chapter 5, the main contribution is testing the proposed two-step method on real world atmospheric turbulence time series. The application to one day of CASES-99 data shows that the method successfully extracts realistic flow structures. Using a number of previous studies that have examined the underlying physical mechanisms of several isolated events in that dataset, we found that the results from our method are consistent with previous studies. The detected events are grouped into six clusters based on their statistical features. It has been shown that the six clusters have very different physical behaviors, although no physical features are used for clustering. Also shown in this chapter is a comparison between the proposed method and the popular wavelet-based event detection method. While wavelet analysis works well when there are relatively well-known structures in time series, it is not good at distinguishing between events and noise of comparable amplitude, and tends to detect structures even when only noise is present in time series since it favors large amplitude events.

In Chapter 6, we analyze ubiquitous flow structures and their effects on the dynamics of stable atmospheric boundary layers. The detected events from the FLOSSII dataset are grouped into three broad clusters, although they span a wide range of different structural shapes, intensities and background conditions. The current classification using clustering is based only on statistical measures of events from time series. Yet, the similarities within and differences between such clusters are associated with corresponding dynamical similarities and differences. For example, the first cluster encompasses smoothest, sometimes wave-like structures that are associated with largest wind speeds, strong turbulence and weak atmospheric stability. The second and third clusters have predominantly sharp structures, which are step-like for the third cluster. They have weak winds and

turbulence, strong stability and are associated with large wind direction shifts. The complexity of structures is thus reduced , which can help in devising their treatment in numerical models.

Appendix A is the manual of the developed R package **TED**, which has been contributed to CRAN (Kang *et al.*, 2014b).

7.2 Future Directions

These results clearly demonstrate that the merits of the which can outperform other investigated methods in the literature. However, the results also raise a number of questions which deserve further research.

1. *Sliding window length.* A subjective choice of the sliding window length when performing the noise tests makes the proposed method flexible to users, which can be chosen according to the time scale users are interested in. However, in order to make the method more convenient to use, further research may involve studying the possibility of an objective suggestion to users, or at least provide some directions, e.g., finding window size ranges within which the method yields the same results.

2. *Feature set.* In the second step of the method, a limited number of generic time series features are used to cluster events. In future work, a more comprehensive set of statistical features could be designed, which calls for an automatic method for choosing useful ones from them. Further, none of the features are related to the nature of atmospheric time series, and future work could investigate more problem specific features, rather than genetic time series metrics.

3. *Overlapping events*. Existence of multi-scale events in atmospheric data and our way of defining events lead to some unavoidable overlaps between events at different scales. This deserves attention in the future.

4. *Event detection from other measured variables in ABL*. The current focus is limited to temperature time series. Further work could study other variables such as wind speed. It would be interesting to see what kind of events can be detected from wind speed and how they are related to those from temperature time series.

5. *Application to other atmospheric turbulence data*. Another natural direction of further research is to apply the method to other datasets with different atmospheric conditions. An interesting question is whether the broad classification of events applies to most of the datasets in ABL.

While this thesis has been focused on event detection and classification in atmospheric turbulence, the developed methods naturally find broader applicability in many other areas that involve the search for patterns in noisy time series. Such applications include financial trading (e.g., Fu *et al.*, 2001), machine condition monitoring (e.g., Chen, 1992), early detection of epidemic outbreaks (e.g., Hashimoto *et al.*, 2000), structure detection in other types of turbulence flows (e.g., Bolzan *et al.*, 2009), to name just a few.

Bibliography

Bibliography

Acevedo OC, Costa FD, Oliveira PES, Puhales FS, Degrazia GA, Roberti DR. 2013. The influence of submeso processes on stable boundary layer similarity relationships. *Journal of the Atmospheric Sciences* **71**(1): 207–225, doi: 10.1175/JAS-D-13-0131.1.

Acharya UR, Chua ECP, Chua KC, Min LC, Tamura T. 2010. Analysis and automatic identification of sleep stages using higher order spectra. *International journal of neural systems* **20**(6): 509–521, doi: 10.1142/S0129065710002589.

Agrawal R, Faloutsos C, Swami AN. 1993. Efficient Similarity Search In Sequence Databases. In: *Proceedings of the 4th International Conference of Foundations of Data Organization and Algorithms (FODO)*, Lomet D (ed). Springer Verlag: Chicago, Illinois, pp. 69–84.

Alahakoon D, Halgamuge S, Srinivasan B. 2000. Dynamic self-organizing maps with controlled growth for knowledge discovery. *IEEE Transactions on Neural Networks* **11**(3): 601–614, doi: 10.1109/72.846732.

Antonia R, Fulachier L. 1989. Topology of a turbulent boundary layer with and without wall suction. *Journal of Fluid Mechanics* **198**: 429–451, doi: 10.1017/S0022112089000200.

Antonia RA, Chambers AJ, Friehe CA, Atta CWV. 1979. Temperature ramps in the atmospheric surface layer. *Journal of the Atmospheric Sciences* **36**(1): 99–108, doi: 10.1175/1520-0469(1979)036<0099:TRITAS>2.0.CO;2.

Arzner K, Knaepen B, Carati D, Denewet N, Vlahos L. 2006. The effect of coherent structures on stochastic acceleration in MHD turbulence. *The Astrophysical Journal* **637**(1): 322–332, doi: 10.1086/498341.

Bai J, Perron P. 2003. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics* **18**(1): 1–22, doi: 10.1002/jae.659.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. Meme suite: tools for motif discovery and searching. *Nucleic Acids Research* **37**(suppl 2): W202–W208, doi: 10.1093/nar/gkp335.

Baklanov AA, Grisogono B, Bornstein R, Mahrt L, Zilitinkevich SS, Taylor P, Larsen SE, Rotach MW, Fernando HJS. 2011. The nature, theory, and modeling of atmospheric planetary boundary layers. *Bulletin of the American Meteorological Society* **92**(2): 123–128, doi: 10.1175/2010BAMS2797.1.

Banerjee A, Dolado JJ, Galbraith JW, Hendry D. 1993. *Co-integration, error correction, and the econometric analysis of non-stationary data*. OUP Catalogue, Oxford University Press, ISBN 9780198288107, doi: 10.1093/0198288107.001.0001.

Banta R, Newsom R, Lundquist J, Pichugina Y, Coulter R, Mahrt L. 2002. Nocturnal low-level jet characteristics over Kansas during CASES-99. *Boundary-Layer Meteorology* **105**(2): 221–252, doi: 10.1023/A:1019992330866.

Barthlott C, Drobinski P, Fesquet C, Dubos T, Pietras C. 2007. Long-term study of coherent structures in the atmospheric surface layer. *Boundary-Layer Meteorology* **125**(1): 1–24, doi: 10.1007/s10546-007-9190-9.

Belušić D, Hrastinski M, Večenaj Ž, Grisogono B. 2013. Wind regimes associated with a mountain gap at the northeastern Adriatic Coast. *Journal of Applied Meteorology and Climatology* **52**(9): 2089–2105, doi: 10.1175/JAMC-D-12-0306.1.

Belušić D, Güttler I. 2010. Can mesoscale motions reproduce meandering motions? *Quarterly Journal of the Royal Meteorological Society* **136**(648): 553–565, doi: 10.1002/qj.606.

Belušić D, Mahrt L. 2008. Estimation of length scales from mesoscale networks. *Tellus A* **60**(4): 706–715, doi: 10.1111/j.1600-0870.2008.00328.x.

Belušić D, Mahrt L. 2012. Is geometry more universal than physics in atmospheric boundary layer flow? *Journal of Geophysical Research* **117**: D09115, doi: 10.1029/2011JD016987.

Benson G, Waterman MS. 1994. A method for fast database search for all k-nucleotide repeats. *Nucleic Acids Research* **22**(22): 4828–4836, doi: 10.1093/nar/22.22.4828.

Bergström H, Högström U. 1989. Turbulent exchange above a pine forest II. organized structures. *Boundary-Layer Meteorology* **49**(3): 231–263, doi: 10.1007/BF00120972.

Bisset D, Antonia R, Browne L. 1990. Spatial organization of large structures in the turbulent far wake of a cylinder. *Journal of Fluid Mechanics* **218**: 439–461, doi: 10.1017/S0022112090001069.

Blackwelder R, Kaplan R. 1976. On the wall structure of the turbulent boundary layer. *Journal of Fluid Mechanics* **76**(1): 89–112, doi: 10.1017/S0022112076003145.

Blumen W, Banta R, Burns SP, Fritts DC, Newsom R, Poulos GS, Sun J. 2001. Turbulence statistics of a kelvin-helmholtz billow event observed in the night-time boundary layer during the cooperative atmosphere-surface exchange study field program. *Dynamics of Atmospheres and Oceans* **34**(2-4): 189–204, doi: 10.1016/S0377-0265(01)00067-7.

Bogard D, Tiederman W. 1986. Burst detection with single-point velocity measurements. *Journal of Fluid Mechanics* **162**: 389–413, doi: 10.1017/S0022112086002094.

Bolzan M, Guarnieri F, Vieira PC. 2009. Comparisons between two wavelet functions in extracting coherent structures from solar wind time series. *Brazilian Journal of Physics* **39**(1): 12–17, doi: 10.1590/S0103-97332009000100002.

Box GEP, Pierce DA. 1970. Distribution of residual autocorrelations in Autoregressive-Integrated moving average time series models. *Journal of the American Statistical Association* **65**(332): 1509–1526, doi: 10.1080/01621459.1970.10481180.

Burge P, Shawe-Taylor J, Cooke C, Moreau Y, Preneel B, Stoermann C. 1997. Fraud detection and management in mobile telecommunications networks. In: *Proceedings of the 1997 European Conference on Security and Detection (ECOS)*. pp. 91–96, doi: 10.1049/cp:19970429.

Campanharo ASLO, Ramos FM, Macau EEN, Rosa RR, Bolzan MJA, Sá LDA. 2008. Searching chaos and coherent structures in the atmospheric turbulence above the Amazon forest. *Philosophical Transactions of the Royal Society A* **366**(1865): 579–589, doi: 10.1098/rsta.2007.2118.

Carpenter GA, Grossberg S. 2010. Adaptive resonance theory. In: *Encyclopedia of Machine Learning*, Sammut C, Webb GI (eds), Springer, ISBN 978-0-387-30768-8, pp. 22–35.

Castro N, Azevedo P. 2010. Multiresolution motif discovery in time series. In: *Proceedings of the 2010 SIAM International Conference on Data Mining (SDM)*. SIAM, pp. 665–676, doi: 10.1137/1.9781611972801.73.

Castro NC, Azevedo PJ. 2012. Significant motifs in time series. *Statistical Analysis and Data Mining* **5**(1): 35–53, doi: 10.1002/sam.11134.

Charrad M, Ghazzali N, Boiteau V, Niknafs A. 2013. NbClust: An examination of indices for determining the number of clusters: NbClust package. URL http://CRAN.R-project.org/package=NbClust. R package version 1.4.

Chen J, Hu F. 2003. Coherent structures detected in atmospheric boundary-layer turbulence using wavelet transforms at Huaihe river basin, China. *Boundary-Layer Meteorology* **107**(2): 429–444, doi: 10.1023/A:1022162030155.

Chen JR. 2005. Making subsequence time series clustering meaningful. In: *Proceedings of the Fifth IEEE International Conference on Data Mining*. IEEE Computer Society, ISBN 0-7695-2278-5, pp. 114–121, doi: 10.1109/ICDM.2005.91.

Chen W, Novak MD, Black TA, Lee X. 1997. Coherent eddies and temperature structure functions for three contrasting surfaces. Part II: Renewal model for sensible heat flux. *Boundary-Layer Meteorology* **84**(1): 125–147, doi: 10.1023/a:1000342918158.

Chen X, Wang M, Zhang Y, Feng Y, Wu Z, Huang NE. 2013. Detecting signal from data with noise: Theory and applications. *Journal of the Atmospheric Sciences* **70**(5): 1489–1504, doi: 10.1175/JAS-D-12-0213.1.

Chen Y. 1992. Machinery condition monitoring by inverse filtering and statistical analysis. *Mechanical Systems and Signal Processing* **6**(2): 177–189, doi: 10.1016/0888-3270(92)90064-P.

Chian ACL, Miranda RA, Koga D, Bolzan MJA, Ramos FM, Rempel EL. 2008. Analysis of phase coherence in fully developed atmospheric turbulence: Amazon forest canopy. *Nonlinear Processes in Geophysics* **15**(4): 567–573, doi: 10.5194/npg-15-567-2008.

Chiu B, Keogh E, Lonardi S. 2003. Probabilistic discovery of time series motifs. In: *Proceedings of the 9th ACM SIGKDD international conference on Knowledge*

discovery and data mining (KDD). ACM: New York, NY, USA, ISBN 1-58113-737-0, pp. 493–498, doi: 10.1145/956750.956808.

Choudhury SM, Shah SL, Thornhill NF. 2008a. Measures of nonlinearity–a review. In: *Diagnosis of Process Nonlinearities and Valve Stiction*, Springer, pp. 69–75.

Choudhury SM, Shah SL, Thornhill NF. 2008b. A nonlinearity measure based on surrogate data analysis. In: *Diagnosis of Process Nonlinearities and Valve Stiction*, Springer, pp. 93–110.

Cobb GW. 1978. The problem of the nile: Conditional solution to a changepoint problem. *Biometrika* **65**(2): 243–251, doi: 10.2307/2335202.

Collineau S, Brunet Y. 1993a. Detection of turbulent coherent motions in a forest canopy part I: Wavelet analysis. *Boundary-Layer Meteorology* **65**(4): 357–379, doi: 10.1007/BF00707033.

Collineau S, Brunet Y. 1993b. Detection of turbulent coherent motions in a forest canopy part II: Time-scales and conditional averages. *Boundary-Layer Meteorology* **66**(1-2): 49–73, doi: 10.1007/BF00705459.

Culotta A. 2010. Towards detecting influenza epidemics by analyzing twitter messages. In: *Proceedings of the First Workshop on Social Media Analytics (SOMA)*. ACM: New York, NY, USA, ISBN 978-1-4503-0217-3, pp. 115–122, doi: 10.1145/1964858.1964874.

Das G, Lin KI, Mannila H, Renganathan G, Smyth P. 1998. Rule discovery from time series. In: *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD)*. AAAI Press: New York, NY, USA, pp. 16–22.

Duda RO, Hart PE, Stork DG. 2001. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2 edn, ISBN 978-0-471-05669-0.

Einaudi F, Bedard AJ, Finnigan JJ. 1989. A climatology of gravity waves and other coherent disturbances at the boulder atmospheric observatory during March–April 1984. *Journal of the Atmospheric Sciences* **46**(3): 303–329, doi: 10.1175/1520-0469(1989)046<0303:ACOGWA>2.0.CO;2.

Einaudi F, Finnigan JJ. 1993. Wave-turbulence dynamics in the stably stratified boundary layer. *Journal of the Atmospheric Sciences* **50**(13): 1841–1864, doi: 10.1175/1520-0469(1993)050<1841:WTDITS>2.0.CO;2.

Elsner JB. 2003. Tracking hurricanes. *Bulletin of the American Meteorological Society* **84**(3): 353–356, doi: 10.1175/BAMS-84-3-353.

Enders W. 2003. *Applied econometric times series*. Wiley, 3 edn, ISBN 978-0-470-57425-6.

Farge M. 1992. Wavelet transforms and their applications to turbulence. *Annual Review of Fluid Mechanics* **24**(1): 395–458.

Feigenwinter C, Vogt R. 2005. Detection and analysis of coherent structures in urban turbulence. *Theoretical and Applied Climatology* **81**(3-4): 219–230, doi: 10.1007/s00704-004-0111-2.

Ferreira PG, Azevedo PJ, Silva CG, Brito RM. 2006. Mining approximate motifs in time series. In: *Discovery Science*, *Lecture Notes in Computer Science*, vol. 4265, Todorovski L, Lavraŷ N, Jantke KP (eds), Springer Berlin Heidelberg, ISBN 978-3-540-46491-4, pp. 89–101, doi: 10.1007/11893318_12.

Finnigan JJ, Einaudi F, Fua D. 1984. The interaction between an internal gravity wave and turbulence in the stably-stratified nocturnal boundary layer. *Journal of the Atmospheric Sciences* **41**(16): 2409–2436, doi: 10.1175/1520-0469(1984)041<2409:TIBAIG>2.0.CO;2.

Firoiu L, Cohen PR. 2002. Segmenting time series with a hybrid neural networks - hidden markov model. In: *Proceedings of the Eighteenth National Conference on*

Artificial Intelligence. American Association for Artificial Intelligence: Menlo Park, CA, USA, ISBN 0-262-51129-0, pp. 247–252.

Fu Tc, Chung Fl, Ng V, Luk R. 2001. Pattern discovery from stock time series using self-organizing maps. In: *Workshop Notes of KDD2001 Workshop on Temporal Data Mining*. pp. 26–29.

Fujimaki R, Yairi T, Machida K. 2005. An approach to spacecraft anomaly detection problem using kernel feature space. In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (KDD)*, Grossman R, Bayardo R, Bennett KP (eds). ACM, pp. 401–410, doi: 10.1145/1081870.1081917.

Ghil M, Allen MR, Dettinger MD, Ide K, Kondrashov D, Mann ME, Robertson AW, Saunders A, Tian Y, Varadi F, Yiou P. 2002. Advanced spectral methods for climatic time series. *Reviews of Geophysics* **40**(1): 1003, doi: 10.1029/2000RG000092.

Gilliam X, Dunyak J, Doggett A, Smith D. 2000. Coherent structure detection using wavelet analysis in long time-series. *Journal of Wind Engineering and Industrial Aerodynamics* **88**(2): 183 – 195, doi: 10.1016/S0167-6105(00)00048-9.

Gluhovsky A, Agee E. 2007. On the analysis of atmospheric and climatic time series. *Journal of Applied Meteorology and Climatology* **46**(7): 1125–1129, doi: 10.1175/JAM2512.1.

Goldin DQ, Mardales R, Nagy G. 2006. In search of meaning for time series subsequence clustering: matching algorithms based on a new distance measure. In: *Proceedings of the 15th ACM international conference on Information and knowledge management*, Yu PS, Tsotras VJ, Fox EA, Liu B (eds). ACM, ISBN 1-59593-433-2, pp. 347–356, doi: 10.1145/1183614.1183666.

Gomes M, Souza A, Guimaraes H, Aguirre L. 2000. Investigation of determinism in heart rate variability. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **10**(2): 398–410, doi: 10.1063/1.166507.

Gray R. 1984. Vector quantization. *ASSP Magazine, IEEE* **1**(2): 4 –29, doi: 10.1109/MASSP.1984.1162229.

Guan, Uberbacher, Guan X, Uberbacher EC. 1996. A fast look-up algorithm for detecting repetitive DNA sequences. In: *Proceedings of the Pacific Symposium on Biocomputing (PSB)*. Singapore, pp. 718–719.

Guarin-Lopez D, Orozco-Gutierrez A, Delgado-Trejos E, Guijarro-Estelles E. 2010. On detecting determinism and nonlinearity in microelectrode recording signals: Approach based on non-stationary surrogate data methods. In: *Proceedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 4032–4035.

Güneş S, Polat K, Yosunkaya Ş. 2010. Efficient sleep stage recognition system based on eeg signal using *k*-means clustering based feature weighting. *Expert Systems with Applications* **37**(12): 7922–7928, doi: 10.1016/j.eswa.2010.04.043.

Hada T, Koga D, Yamamoto E. 2003. Phase coherence of MHD waves in the solar wind. *Space science reviews* **107**(1-2): 463–466, doi: 10.1023/A:1025506124402.

Harvey AC, Durbin J. 1986. The effects of seat belt legislation on british road casualties: A case study in structural time series modelling. *International Journal of Forecasting* **2**(4): 496–497, doi: 10.1016/0169-2070(86)90097-X.

Hashimoto S, Murakami Y, Taniguchi K, Nagai M. 2000. Detection of epidemics in their early stage through infectious disease surveillance. *International journal of epidemiology* **29**(5): 905–910, doi: 10.1093/ije/29.5.905.

Hochheiser H, Shneiderman B. 2002. Visual queries for finding patterns in time series data. Technical report, Computer Science Department, University of Maryland.

Hohreiter V. 2008. Finescale structure and dynamics of an atmospheric temperature interface. *Journal of the Atmospheric Sciences* **65**(5): 1701–1710, doi: 10.1175/2007JAS2576.1.

Holtslag AAM, Svensson G, Baas P, Basu S, Beare B, Beljaars ACM, Bosveld FC, Cuxart J, Lindvall J, Steeneveld GJ, Tjernström M, Van De Wiel BJH. 2013. Stable atmospheric boundary layers and diurnal cycles: Challenges for weather and climate models. *Bulletin of the American Meteorological Society* **94**(11): 1691–1706, doi: 10.1175/BAMS-D-11-00187.1.

Hudgins L, Kaspersen JH. 1999. Wavelets and detection of coherent structures in fluid turbulence. In: *Wavelets in Physics*, vol. 1. pp. 201–226.

Hussain AKMF. 1981. Role of coherent structures in turbulent shear flows. *Proceedings of the Indian Academy of Sciences Section C: Engineering Sciences* **4**(2): 129–175, doi: 10.1007/BF02896739.

Hussain AKMF. 1983. Coherent structures – reality and myth. *Physics of Fluids* **26**(10): 2816–2850, doi: 10.1063/1.864048.

Hussain AKMF. 1986. Coherent structures and turbulence. *Journal of Fluid Mechanics* **173**: 303–356, doi: 10.1017/S0022112086001192.

Jeong J, Gore JC, Peterson BS. 2002. Detecting determinism in short time series, with an application to the analysis of a stationary EEG recording. *Biological cybernetics* **86**(5): 335–342, doi: 10.1007/s00422-001-0299-5.

Jirayusakul A, Auwatanamongkol S. 2007. A supervised growing neural gas algorithm for cluster analysis. *International Journal of Hybrid Intelligent Systems* **4**(2): 129–141.

Jonsson P, Eklundh L. 2002. Seasonality extraction by function fitting to time-series of satellite sensor data. *IEEE Transactions on Geoscience and Remote Sensing* **40**(8): 1824 – 1832, doi: 10.1109/TGRS.2002.802519.

Jouini J, Boutahar M. 2005. Evidence on structural changes in U.S. time series. *Economic Modelling* **22**(3): 391–422, doi: 10.1016/j.econmod.2004.06.003.

Kang Y. 2012. Real-time change detection in time series based on growing feature quantization. In: *Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–6, doi: 10.1109/IJCNN.2012.6252381.

Kang Y, Belušić D, Smith-Miles K. 2014a. Classes of structures in the stable atmospheric boundary layer (*Submitted*). *Quarterly Journal of the Royal Meteorological Society* .

Kang Y, Belušić D, Smith-Miles K. 2014b. **TED**: Turbulence Event Detection and classification. URL `http://CRAN.R-project.org/package=TED`. R package version 1.0.

Kang Y, Belušić D, Smith-Miles K. 2014c. Detecting and classifying events in noisy time series. *Journal of the Atmospheric Sciences* **71**(3): 1090–1104, doi: 10.1175/JAS-D-13-0182.1.

Kang Y, Belušić D, Smith-Miles K. 2014d. A note on the relationship between turbulent coherent structures and phase correlation. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **24**(2): 023114, doi: http://dx.doi.org/10.1063/1.4875260.

Kang Y, Smith-Miles K, Belušić D. 2013. How to extract meaningful shapes from noisy time-series subsequences? In: *Proceedings of the 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE, pp. 65–72, doi: 10.1109/CIDM.2013.6597219.

Katul G, Kuhn G, Schieldge J, Hsieh CI. 1997. The ejection-sweep character of scalar fluxes in the unstable surface layer. *Boundary-Layer Meteorology* **83**(1): 1–26, doi: 10.1023/A:1000293516830.

Kawahara Y, Sugiyama M. 2009. Change-point detection in time-series data by direct density-ratio estimation. In: *Proceedings of the 2009 SIAM International Conference on Data Mining (ICDM)*. SIAM, pp. 389–400, doi: 10.1137/1.9781611972795.34.

Keogh E, Chakrabarti K, Pazzani M, Mehrotra S. 2001a. Locally adaptive dimensionality reduction for indexing large time series databases. *ACM SIGMOD Record* **30**(2): 151–162, doi: 10.1145/376284.375680.

Keogh E, Chakrabarti K, Pazzani MJ, Mehrotra S. 2000. Dimensionality reduction for fast similarity search in large time series databases. *Journal of Knowledge and Information Systems* **3**(3): 263–286.

Keogh E, Chu S, Hart D, Pazzani MJ. 2001b. An online algorithm for segmenting time series. In: *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM)*. IEEE Computer Society: Washington, DC, USA, ISBN 0-7695-1119-8, pp. 289–296, doi: 10.1109/ICDM.2001.989531.

Keogh E, Hochheiser H, Shneiderman B. 2002. An augmented visual query mechanism for finding patterns in time series data. In: *Flexible Query Answering Systems*, *Lecture Notes in Computer Science*, vol. 2522, Carbonell J, Siekmann J, Andreasen T, Christiansen H, Motro A, Larsen H (eds), Springer Berlin Heidelberg, ISBN 978-3-540-00074-7, pp. 240–250, doi: 10.1007/3-540-36109-X_19.

Keogh E, Kasetty S. 2002. On the need for time series data mining benchmarks: a survey and empirical demonstration. In: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM: Edmonton, Alberta, Canada, pp. 102–111.

Keogh E, Lin J, Truppel W. 2003. Clustering of time series subsequences is meaningless: implications for previous and future research. In: *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM)*. pp. 115–122, doi: 10.1109/ICDM.2003.1250910.

Kifer D, Ben-David S, Gehrke J. 2004. Detecting change in data streams. In: *Proceedings of the Thirtieth International Conference on Very Large Data Bases (VLDB)*, vol. 30. VLDB Endowment, ISBN 0-12-088469-0, pp. 180–191, doi: 10.1007/PL00011669.

Koga D, Chian ACL, Hada T, Rempel EL. 2008. Experimental evidence of phase coherence of magnetohydrodynamic turbulence in the solar wind: GEOTAIL satellite data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **366**: 447–457, doi: 10.1098/rsta.2007.2102.

Kugiumtzis D. 2002. Surrogate data test on time series. In: *Modelling and Forecasting Financial Data*, Springer, pp. 267–282.

Kumar P, Foufoula-Georgiou E. 1994. Wavelet analysis in geophysics: An introduction. In: *Wavelets in geophysics*, Foufoula-Georgiou E, Kumar P (eds), Academic Press New York, pp. 1–43.

Kuznetsov E, Zakharov V. 2000. Nonlinear coherent phenomena in continuous media. In: *Nonlinear Science at the Dawn of the 21st Century*, Christiansen P, Sorensen M, Scott A (eds), Springer-Verlag, Berlin, pp. 3–45.

Lam HT, Calders T, Pham N. 2011. Online discovery of top-k similar motifs in time series data. In: *Proceedings of SIAM International Conference on Data Mining (SDM)*. SIAM / Omnipress: Mesa, Arizona, USA, pp. 1004–1015.

Lan LW, Lin FY, Kuo AY. 2005. Identification for chaotic phenomena in short-term traffic flows: A parsimony procedure with surrogate data. *Journal of the Eastern Asia Society for Transportation Studies* **6**: 1518–1533.

Lee YH, Chen YS, Chen LF. 2009. Automated sleep staging using single EEG channel for REM sleep deprivation. In: *Ninth IEEE International Conference on Bioinformatics and BioEngineering (BIBE)*. pp. 439–442, doi: 10.1109/BIBE.2009.68.

Lin FY. 2005. Traffic flow analysis with diffeerent time scales. *Journal of the Eastern Asia Society for Transportation Studies* **6**: 1624–1636.

Lin J, Keogh E, Lonardi S. 2005. Visualizing and discovering non-trivial patterns in large time series databases. *Information Visualization* **4**(2): 61–82, doi: 10.1057/palgrave.ivs.9500089.

Lin J, Keogh E, Lonardi S, Chiu B. 2003. A symbolic representation of time series, with implications for streaming algorithms. In: *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD)*. ACM: New York, NY, USA, pp. 2–11, doi: 10.1145/882082.882086.

Lin J, Keogh E, Lonardi S, Patel P. 2002. Finding Motifs in Time Series. In: *Proceedings of the 2nd Workshop on Temporal Data Mining, at the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM: Edmonton, Alberta, Canada, pp. 53–68, doi: 10.1.1.19.6629.

Lughofer E. 2008. Extensions of vector quantization for incremental clustering. *Pattern Recognition* **41**(3): 995–1011, doi: 10.1016/j.patcog.2007.07.019.

Lunetta RS, Knight JF, Ediriwickrema J, Lyon JG, Worthy LD. 2006. Land-cover change detection using multi-temporal MODIS NDVI data. *Remote Sensing of Environment* **105**(2): 142–154, doi: 10.1016/j.rse.2006.06.018.

Mahrt L. 1991. Eddy asymmetry in the sheared heated boundary layer. *Journal of the Atmospheric Sciences* **48**(3): 472–492, doi: 10.1175/1520-0469(1991)048<0472:EAITSH>2.0.CO;2.

Mahrt L. 1999. Stratified atmospheric boundary layers. *Boundary-Layer Meteorology* **90**(3): 375–396, doi: 10.1023/A:1001765727956.

Mahrt L. 2009. Characteristics of submeso winds in the stable boundary layer. *Boundary-Layer Meteorology* **130**(1): 1–14, doi: 10.1007/s10546-008-9336-4.

Mahrt L. 2010. Common microfronts and other solitary events in the nocturnal boundary layer. *Quarterly Journal of the Royal Meteorological Society* **136**(652): 1712–1722, doi: 10.1002/qj.694.

Mahrt L. 2011a. Surface wind direction variability. *Journal of Applied Meteorology and Climatology* **50**(1): 144–152, doi: 10.1175/2010JAMC2560.1.

Mahrt L. 2011b. Surface wind direction variability. *Journal of Applied Meteorology and Climatology* **50**(1): 144–152, doi: 10.1175/2010JAMC2560.1.

Mahrt L. 2014. Stably stratified atmospheric boundary layers. *Annual Review of Fluid Mechanics* **46**(1): 23–45, doi: 10.1146/annurev-fluid-010313-141354.

Mahrt L, Gibson W. 1992. Flux decomposition into coherent structures. *Boundary-Layer Meteorology* **60**(1-2): 143–168, doi: 10.1007/BF00122065.

Mahrt L, Richardson S, Seaman N, Stauffer D. 2012. Turbulence in the nocturnal boundary layer with light and variable winds. *Quarterly Journal of the Royal Meteorological Society* **138**(667): 1430–1439, doi: 10.1002/qj.1884.

Mahrt L, Thomas C, Richardson S, Seaman N, Stauffer D, Zeeman M. 2013. Non-stationary generation of weak turbulence for very stable and weak-wind conditions. *Boundary-Layer Meteorology* **147**(2): 179–199, doi: 10.1007/s10546-012-9782-x.

Maiwald T, Mammen E, Nandi S, Timmer J. 2008. Surrogate data – a qualitative and quantitative analysis. In: *Mathematical Methods in Signal Processing and Digital Image Analysis*, Dahlhaus R, Kurths J, Maass P, Timmer J (eds), Understanding

Complex Systems, Springer Berlin Heidelberg, ISBN 978-3-540-75631-6, pp. 41–74, doi: 10.1007/978-3-540-75632-3_2.

McGovern A, Rosendahl DH, Brown RA, Droegemeier KK. 2011. Identifying predictive multi-dimensional time series motifs: an application to severe weather prediction. *Data Mining and Knowledge Discovery* **22**(1-2): 232–258, doi: 10.1007/s10618-010-0193-7.

Miranda RA, Rempel EL, Chian ACL, Seehafer N, Toledo BA, Muñoz PR. 2013. Lagrangian coherent structures at the onset of hyperchaos in the two-dimensional navier-stokes equations. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **23**(3): 033 107–033 107, doi: 10.1063/1.4811297.

Mueen A, Keogh E. 2010. Online discovery and maintenance of time series motifs. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM: New York, NY, USA, ISBN 978-1-4503-0055-1, pp. 1089–1098, doi: 10.1145/1835804.1835941.

Mueen A, Keogh E, Bigdely-Shamlo N. 2009a. Finding time series motifs in disk-resident data. In: *Proceedings of the 9th IEEE International Conference on Data Mining (ICDM)*. IEEE Computer Society: Washington, DC, USA, ISBN 978-0-7695-3895-2, pp. 367–376, doi: 10.1109/ICDM.2009.15.

Mueen A, Keogh E, Zhu Q, Cash S, Westover MB. 2009b. Exact discovery of time series motifs. In: *Proceedings of the 2009 SIAM International Conference on Data Mining (SDM)*. SIAM: Sparks, Nevada, USA, pp. 473–484.

Nappo C, Sun J, Mahrt L, Belušić D. 2014. Determining wave–turbulence interactions in the stable boundary layer. *Bulletin of the American Meteorological Society* **95**(1): ES11–ES13, doi: 10.1175/BAMS-D-12-00235.1.

Ouellette NT. 2012. On the dynamical role of coherent structures in turbulence. *Comptes Rendus Physique* **13**(9): 866–877, doi: 10.1016/j.crhy.2012.09.006.

Peacock T, Dabiri J. 2010. Introduction to focus issue: Lagrangian coherent structures. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **20**(1): 017 501–017 501, doi: 10.1063/1.3278173.

Percival D. 2010. An ominibus test for red noise, with applications to climatology time series. In: *An Ominibus Test for Red Noise, with Applications to Climatology Time Series*. Presented 26 March 2010 at CSIRO Mathematics, Informatics and Statistics's TechFest: Modelling Complex Systems, Hobart, Tasmania.

Pereda E, Gamundi A, Rial R, Gonzalez J. 1998. Non-linear behaviour of human eeg: fractal exponent versus correlation dimension in awake and sleep stages. *Neuroscience letters* **250**(2): 91–94, doi: 10.1016/S0304-3940(98)00435-2.

Perron P. 1988. Trends and random walks in macroeconomic time series: Further evidence from a new approach. *Journal of economic dynamics and control* **12**(2-3): 297–332, doi: 10.1016/0165-1889(88)90043-7.

Pong Chan K, Fu AWC. 1999. Efficient time series matching by wavelets. In: *Proceedings of the 15th International Conference on Data Engineering (ICDE)*, Kitsuregawa M, Papazoglou MP, Pu C (eds). IEEE Computer Society: Sydney, Australia, ISBN 0-7695-0071-4, pp. 126–133, doi: 10.1109/ICDE.1999.754915.

Pope M, Jakob C, Reeder MJ. 2009. Objective classification of tropical mesoscale convective systems. *Journal of Climate* **22**(22): 5797–5808, doi: 10.1175/2009JCLI2777.1.

Poulos GS, Blumen W, Fritts DC, Lundquist JK, Sun J, Burns SP, Nappo C, Banta R, Newsom R, Cuxart J, Terradellas E, Balsley B, Jensen M. 2002. CASES-99: A comprehensive investigation of the stable nocturnal boundary layer. *Bulletin of the American Meteorological Society* **83**: 555–581, doi: 10.1175/1520-0477(2002)083<0555:CACIOT>2.3.CO;2.

Provenzale A, Smith L, Vio R, Murante G. 1992. Distinguishing between low-dimensional dynamics and randomness in measured time series. *Physica D: Nonlinear Phenomena* **58**(1): 31–49, doi: 10.1016/0167-2789(92)90100-2.

R Core Team. 2013. R: A language and environment for statistical computing. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

Rajagopalan S, Antonia R. 1982. Use of a quadrant analysis technique to identify coherent structures in a turbulent boundary layer. *Physics of Fluids (1958-1988)* **25**(6): 949–956, doi: 10.1063/1.863848.

Rhodes C, Morari M. 1998. Determining the model order of nonlinear input/output systems. *AIChE Journal* **44**(1): 151–163, doi: 10.1002/aic.690440116.

Rigoutsos I, Floratos A. 1998. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics* **14**(1): 55–67, doi: 10.1093/bioinformatics/14.1.55.

Robinson SK. 1991. Coherent motions in the turbulent boundary layer. *Annual Review of Fluid Mechanics* **23**(1): 601–639, doi: 10.1146/annurev.fl.24.010192.002143.

Sahraoui F. 2008. Diagnosis of magnetic structures and intermittency in space-plasma turbulence using the technique of surrogate data. *Physical Review E* **78**: 026 402, doi: 10.1103/PhysRevE.78.026402.

Salmon B, Olivier J, Wessels K, Kleynhans W, van den Bergh F, Steenkamp K. 2011. Unsupervised land cover change detection: Meaningful sequential time series analysis. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of* **4**(2): 327 –335, doi: 10.1109/JSTARS.2010.2053918.

Sandu I, Beljaars A, Bechtold P, Mauritsen T, Balsamo G. 2013. Why is it so difficult to represent stably stratified conditions in numerical weather prediction (NWP)

models? *Journal of Advances in Modeling Earth Systems* **5**(2): 117–133, doi: 10.1002/jame.20013.

Schreiber T, Schmitz A. 1997. Discrimination power of measures for nonlinearity in a time series. *Physical Review E* **55**(5): 5443, doi: 10.1103/PhysRevE.55.5443.

Schreiber T, Schmitz A. 2000. Surrogate time series. *Physica D: Nonlinear Phenomena* **142**(3): 346–382, doi: 10.1016/S0167-2789(00)00043-9.

Segalini A, Alfredsson P. 2012. Techniques for the eduction of coherent structures from flow measurements in the atmospheric boundary layer. *Boundary-Layer Meteorology* **143**(3): 433–450, doi: 10.1007/s10546-012-9708-7.

Shapland T, McElrone A, Snyder R, Paw U K. 2012a. Structure function analysis of two-scale scalar ramps. Part I: Theory and modelling. *Boundary-Layer Meteorology* **145**(1): 5–25, doi: 10.1007/s10546-012-9742-5.

Shapland T, McElrone A, Snyder R, Paw U K. 2012b. Structure function analysis of two-scale scalar ramps. Part II: Ramp characteristics and surface renewal flux estimation. *Boundary-Layer Meteorology* **145**(1): 27–44, doi: 10.1007/s10546-012-9740-7.

Sharifzadeh M, Azmoodeh F, Shahabi C. 2005. Change detection in time series data using wavelet footprints. In: *Advances in Spatial and Temporal Databases*, *Lecture Notes in Computer Science*, vol. 3633, Bauzer Medeiros C, Egenhofer M, Bertino E (eds). Springer Berlin Heidelberg, ISBN 978-3-540-28127-6, pp. 127–144, doi: 10.1007/11535331_8.

Shieh J, Keogh E. 2008. iSAX: Indexing and mining terabyte sized time series. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM: New York, NY, USA, pp. 623–631, doi: 10.1145/1401890.1401966.

Singh S. 2000. Pattern modelling in time-series forecasting. *Cybernetics and Systems - an International Journal* **31**(1): 49–65, doi: 10.1080/019697200124919.

Sledge I, Keller J. 2008. Growing neural gas for temporal clustering. In: *Proceedings of the 19th International Conference on Pattern Recognition (ICPR)*. pp. 1–4, doi: 10.1109/ICPR.2008.4761768.

Stanus E. 1986. Computerized sleep stages analysis. *Signal Processing* **10**(1): 101–102, doi: 10.1016/0165-1684(86)90070-8.

Steiner A, Pressley S, Botros A, Jones E, Chung S, Edburg S. 2011. Analysis of coherent structures and atmosphere-canopy coupling strength during the CABINEX field campaign. *Atmospheric Chemistry and Physics* **11**(23): 11 921–11 936, doi: 10.5194/acp-11-11921-2011.

Storch HV, Zwiers FW. 1999. *Statistical analysis in climate research*. Cambridge University Press, ISBN 978-0521012300.

Sugihara G, May RM. 1990. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature* **344**(6268): 734–741, doi: doi:10.1038/344734a0.

Sun J, Burns SP, Lenschow DH, Banta R, Newsom R, Coulter R, Frasier S, Ince T, Nappo C, Cuxart J, Blumen W, Lee X, Hu XZ. 2002. Intermittent turbulence associated with a density current passage in the stable boundary layer. *Boundary-Layer Meteorology* **105**(2): 199–219, doi: 10.1023/A:1019969131774.

Sun J, Lenschow DH, Burns SP, Banta RM, Newsom RK, Coulter R, Frasier S, Ince T, Nappo C, Balsley B, Jensen M, Mahrt L, Miller D, Skelly B. 2004. Atmospheric disturbances that generate intermittent turbulence in nocturnal boundary layers. *Boundary-Layer Meteorology* **110**(2): 255–279, doi: 10.1023/A:1026097926169.

Sun J, Mahrt L, Banta RM, Pichugina YL. 2012. Turbulence regimes and turbulence intermittency in the stable boundary layer during CASES-99. *Journal of the Atmospheric Sciences* **69**(1): 338–351, doi: 10.1175/JAS-D-11-082.1.

Tanaka Y, Iwamoto K, Uehara K. 2005. Discovery of time-series motif from multi-dimensional data based on mdl principle. *Machine Learning* **58**(2-3): 269–300, doi: 10.1007/s10994-005-5829-2.

Tang W, Chan PW, Haller G. 2010. Accurate extraction of lagrangian coherent structures over finite domains with application to flight data analysis over Hong Kong international airport. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **20**(1): 017 502–017 502, doi: 10.1063/1.3276061.

Tang W, Peacock T. 2010. Lagrangian coherent structures and internal wave attractors. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **20**(1): 017 508–017 508, doi: 10.1063/1.3273054.

Theiler J, Eubank S, Longtin A, Galdrikian B, Doyne Farmer J. 1992. Testing for nonlinearity in time series: the method of surrogate data. *Physica D: Nonlinear Phenomena* **58**(1): 77–94, doi: 10.1016/0167-2789(92)90102-S.

Thomas C, Foken T. 2005. Detection of long-term coherent exchange over spruce forest using wavelet analysis. *Theoretical and Applied Climatology* **80**(2-4): 91–104, doi: 10.1007/s00704-004-0093-0.

Thomas C, Foken T. 2007a. Flux contribution of coherent structures and its implications for the exchange of energy and matter in a tall spruce canopy. *Boundary-Layer Meteorology* **123**(2): 317–337, doi: 10.1007/s10546-006-9144-7.

Thomas C, Foken T. 2007b. Organised motion in a tall spruce canopy: temporal scales, structure spacing and terrain effects. *Boundary-Layer Meteorology* **122**(1): 123–147, doi: 10.1007/s10546-006-9087-z.

Thomas C, Mayer JC, Meixner FX, Foken T. 2006. Analysis of low-frequency turbulence above tall vegetation using a doppler sodar. *Boundary-Layer Meteorology* **119**(3): 563–587, doi: 10.1007/s10546-005-9038-0.

Tompa M. 1999. An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. In: *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, ISBN 1-57735-083-9, pp. 262–271.

Tsai CY, Shieh YC. 2009. A change detection method for sequential patterns. *Decision Support Systems* **46**(2): 501–511, doi: 10.1016/j.dss.2008.09.003.

Tsay RS. 2005. *Analysis of financial time series*. Wiley-Interscience, 2 edn, ISBN 978-1-118-01709-8.

Turner B, Leclerc M. 1994. Conditional sampling of coherent structures in atmospheric turbulence using the wavelet transform. *Journal of Atmospheric and Oceanic Technology* **11**(1): 205–209.

Verbesselt J, Hyndman R, Newnham G, Culvenor D. 2010a. Detecting trend and seasonal changes in satellite image time series. *Remote Sensing of Environment* **114**(1): 106–115, doi: DOI:10.1016/j.rse.2009.08.014.

Verbesselt J, Hyndman R, Zeileis A, Culvenor D. 2010b. Phenological change detection while accounting for abrupt and gradual trends in satellite image time series. *Remote Sensing of Environment* **114**(12): 2970–2980, doi: 10.1016/j.rse.2010.08.003.

Viana S, Terradellas S, Yagüe C. 2010. Analysis of gravity waves generated at the top of a drainage flow. *Journal of the Atmospheric Sciences* **67**(12): 3949–3966, doi: 10.1175/2010JAS3508.1.

Viana S, Yagüe C, Maqueda G. 2009. Propagation and effects of a mesoscale gravity wave over a weakly-stratified nocturnal boundary layer during the

SABLES2006 field campaign. *Boundary-Layer Meteorology* **133**(2): 165–188, doi: 10.1007/s10546-009-9420-4.

Vickers D, Mahrt L. 2007. Observations of the cross-wind velocity variance in the stable boundary layer. *Environmental Fluid Mechanics* **7**(1): 55–71, doi: 10.1007/s10652-006-9010-7.

Wallace JM, Eckelmann H, Brodkey RS. 1972. The wall region in turbulent shear flow. *Journal of Fluid Mechanics* **54**(01): 39–48, doi: 10.1017/S0022112072000515.

Wang X, Smith KA, Hyndman RJ. 2006. Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery* **13**(3): 335–364, doi: 10.1007/s10618-005-0039-x.

Waser M. 2010. Nonlinear dependencies in and between time series. Master's thesis, Vienna University of Technology, Vienna.

Weber RO, Kaufmann P. 1995. Automated classification scheme for wind fields. *Journal of Applied Meteorology* **34**(5): 1133–1141, doi: 10.1175/1520-0450(1995)034<1133:ACSFWF>2.0.CO;2.

Wilczak JM. 1984. Large-scale eddies in the unstably stratified atmospheric surface layer. Part I: Velocity and temperature structure. *Journal of the Atmospheric Sciences* **41**(24): 3537–3550, doi: 10.1175/1520-0469(1984)041<3537:LSEITU>2.0.CO;2.

Williams A, Hacker J. 1992. The composite shape and structure of coherent eddies in the convective boundary layer. *Boundary-Layer Meteorology* **61**(3): 213–245, doi: 10.1007/BF02042933.

Williams G. 2011. *Data mining with Rattle and R: The art of excavating data for knowledge discovery (use R!)*. Springer, doi: 10.1007/978-1-4419-9890-3.

Wilson W, Birkin P, Aickelin U. 2008. The motif tracking algorithm. *International Journal of Automation and Computing* **5**(1): 32–44, doi: 10.1007/s11633-008-0032-0.

Yankov D, Keogh E, Medina J, Chiu B, Zordan V. 2007. Detecting time series motifs under uniform scaling. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07. ACM: New York, NY, USA, ISBN 978-1-59593-609-7, pp. 844–853, doi: 10.1145/1281192.1281282.

Zeileis A, Kleiber C, Krämer W, Hornik K. 2003. Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis* **44**(1-2): 109–123, doi: 0.1016/S0167-9473(03)00030-6.

Zhu W, van Hout R, Katz J. 2007. PIV measurements in the atmospheric boundary layer within and above a mature corn canopy. part ii: Quadrant-hole analysis. *Journal of the atmospheric sciences* **64**(8): 2825–2838, doi: 10.1175/JAS3990.1.

Zivot E, Andrews DWK. 1992. Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of Business & Economic Statistics* **20**(1): 25–44, doi: 10.1198/073500102753410372.

Appendix

Appendix A

Manual for the R package TED: Turbulence Event Detection and classification

Package 'TED'

Type Package

Title Turbulence time series Event Detection and classification

Version 1.0

Date 2014-05-08

Author Yanfei Kang, Danijel Belusic and Kate Smith-Miles

Maintainer Yanfei Kang <yanfei.kang@monash.edu>

Description

TED performs event detection and classification in turbulence time series.

LazyLoad yes

Repository CRAN

Depends R (>= 3.0.2)

Imports foreach, zoo, fields, animation, geoR, tcltk, utils, RcppArmadillo

Suggests doMC

NeedsCompilation no

License GPL (>=2)

R topics documented:

`ted-package` *Detect and classify events from turbulence time series*

Description

TED performs event detection and classification in turbulence time series. The method consists of two steps. The event detection step locates and detects events by performing noise tests on sliding subsequences extracted from the time series. A subsequence is considered to be a potential event if its characteristics are significantly different from noise. The event is defined only if the consecutive sequence of potential events is long enough. This step does not rely on pre-assumption of events in terms of their magnitude, geometry, or stationarity. The main function `eventDetection` should be used for this step. The event classification step is to classify the events into groups with similar global characteristics. Each event is summarised using a feature vector,

and then the events are clustered according to the Euclidean distances among the feature vectors. The main function `eventCluster` should be used for the classification step. Examples of event detection and classification can be found in the package for both artificial data and real world turbulence data.

Details

The package contains two main functions:

`eventDetection`: to detect events from time series as described in Kang et al. (2014b).

`eventCluster`: to classify the detect events from time series as described in Kang et al. (2014b).

The package also contains functions for visualising the events:

`plotevents`: to plot the detected and classified events.

`aniplotevents`: to generate a gif to visualize the event detection process.

Other sub-functions are:

`cbfs`: to generate an artificial event with white noise.

`cbfs_red`: to generate an artificial event with red noise.

`detrendc`: to conditionally detrend a time series.

`eventExtraction`: to extract events from the noise test results of a time series.

`measures`: to calculate statistical characteristics of an event.

`noiseTests`: to perform noise tests for a time series.

`ts2mat`: to reshape a vector into a matrix.

`ur.za.fast`: unit root test for events considering a structrual break.

The real world turbulence dataset used in this package is available by loading:

`CASES99`: one day of 1-s averages of the thermocouple temperature data from CASES-99 dataset (Poulos et al. (2002)).

Author(s)

Yanfei Kang, Danijel Belusic and Kate Smith-Miles

Maintainer: Yanfei Kang <yanfei.kang@monash.edu>

References

Yanfei Kang, Kate Smith-Miles, Danijel Belusic (2013). How to extract meaningful shapes from noisy time-series subsequences? *2013 IEEE Symposium on Computational Intelligence and Data Mining*, Singapore, 65-72. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6597219&isnumber=6597208.

Yanfei Kang, Danijel Belusic, Kate Smith-Miles (2014a). Detecting and Classifying Events in Noisy Time Series. *J. Atmos. Sci.*, 71, 1090-1104. http://dx.doi.org/10.1175/JAS-D-13-0182.1.

Yanfei Kang, Danijel Belusic, Kate Smith-Miles (2014b). Classes of structures in the stable atmospheric boundary layer. Submitted to Quarterly Journal of the Royal Meteorological Society.

Xiaozhe Wang, Kate Smith-Miles and Rob Hyndman (2005). Characteristic-Based Clustering for Time Series Data. *Data Mining and Knowledge Discovery*. 13(3), 335-364. http://dx.doi.org//10.1007/s10618-005-0039-x.

Gregory S. Poulos, William Blumen, David C. Fritts, Julie K. Lundquist, Jielun Sun, Sean P. Burns, Carmen Nappo, Robert Banta, Rob Newsom, Joan Cuxart, Enric Terradellas, Ben Balsley, and Michael Jensen. CASES-99: A comprehensive investigation of the stable nocturnal boundary layer (2002). *Bulletin of the American Meteorological Society*, 83(4):555-581.

`aniplotevents` *Generate a gif to visualise the event detection process*

Description

This function generates a gif file demonstrating how the event detection process is implemented.

Usage

```
aniplotevents(x, w, noiseType = c("white", "red"), alpha = 0.05,
  main = "Animation plot of events", xlab = "t", ylab = "x",
  movie.name = "animation.gif", interval = 0.05,
  ani.width = 1000,  ani.height = 400, outdir = getwd())
```

Arguments

`x`	a vector or a time series.
`w`	a scalar specifying the size of the sliding window.
`noiseType`	background noise type assumed for x. There are two options: white noise or red noise.
`alpha`	the significance level. When the noise test p value of the subsequence is smaller than this significance level, it is defined as a potential event.
`main`	title of the animiation plot; default is 'Animation plot of event detection'.
`xlab`	x label of the animation plot; default is 't'.
`ylab`	y label of the animation plot; default is 'x'.

`movie.name`	name of the output gif file; default is 'animation.gif'.
`interval`	a positive number to set the time interval of the animation (unit in seconds); default is 0.05.
`ani.width`	width of the gif file (unit in px), default is 1000.
`ani.height`	height of the gif file (unit in px); default is 400.
`outdir`	character: specify the output directory when exporting the animations; default to be the current working directory.

Value

...

References

Yihui Xie (2013). Animation: An R Package for Creating Animations and Demonstrating Statistical Methods. *Journal of Statistical Software*, 53(1), 1-27. `http://www.jstatsoft.org/v53/i01/`.

See Also

`noiseTests`, `eventExtraction`, `plotevents`

Examples

```
set.seed(123)
# generate an artificial time series
x=c(rnorm(128),cbfs(type="box"),rnorm(128),cbfs(type="rc"),rnorm(128))
# generate a gif file to show the event detection process
# aniplotevents(x,w=128,noiseType="white",outdir=getwd())
```

`CASES99`	*One day of 1-s averages of the thermocouple temperature data from CASES-99 dataset*

Description

These are 1-s averages of the CASES-99 (Poulos et al. 2002) thermocouple temperature data at the seventh level (9.5 m) from 1100 LST 5 October to 1100 LST 6 October.

Usage

```
data(CASES99)
```

Details

Cooperative Atmosphere-Surface Exchange Study (CASES-99) was conducted over a relatively flat-terrain rural grassland site near Leon, Kansas, during October 1999. As a part of the extensive observations, a 60-m tower was equipped with thermocouples at 34 vertical levels (0.23, 0.63, 2.3 m, and every 1.8 m above 2.3 m) that sampled air temperature five times per second (Sun et al. 2012), while 20-Hz sonic anemometer measurements were taken at seven levels (1.5, 5, 10, 20, 30, 40, 50, and 55 m). 1-s averages of the CASES-99 thermocouple temperature data at the seventh level (9.5 m) from 1100 LST 5 October to 1100 LST 6 October are taken as an example for detection and clustering of events.

Source

Gregory S. Poulos, William Blumen, David C. Fritts, Julie K. Lundquist, Jielun Sun, Sean P. Burns, Carmen Nappo, Robert Banta, Rob Newsom, Joan Cuxart,

Enric Terradellas, Ben Balsley, and Michael Jensen. CASES-99: A comprehensive investigation of the stable nocturnal boundary layer (2002). *Bulletin of the American Meteorological Society*, 83(4):555-581.

Examples

```
data(CASES99)
```

`cbfs` *Generate an artificial event with white noise*

Description

This function generates a box, cliff-ramp, ramp-cliff or a sine function with different levels of white noise as the background noise. Length of the generated event is 128. Generation of events are similar to that of Cylinder-Bell-Funnel dataset in the reference below (Keogh and Lin 2005).

Usage

```
cbfs(type = c("box", "rc", "cr", "sine"), A = 10, sigma = 1)
```

Arguments

`type`	type of the event to be generated. There are four options: 'box', 'rc','cr','sine' representing a box, cliff-ramp, ramp-cliff or a sine function.
`A`	amplitude of the event; default is 10.
`sigma`	a scalar specifying the level of white noise. Default is 1, which means the standard deviation of noise is 1.

Value

an artificial event with white noise.

References

Eamonn Keogh and Jessica Lin (2005). Clustering of time-series subsequences is meaningless: implications for previous and future research. *Knowl. Inf. Syst.*, 8(2), 154-177. http://dblp.uni-trier.de/db/journals/kais/kais8.html#KeoghL05.

Yanfei Kang, Kate Smith-Miles, Danijel Belusic (2013). How to extract meaningful shapes from noisy time-series subsequences? *2013 IEEE Symposium on Computational Intelligence and Data Mining*, Singapore, 65-72. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6597219&isnumber=6597208.

Yanfei Kang, Danijel Belusic, Kate Smith-Miles (2014). Detecting and Classifying Events in Noisy Time Series. *J. Atmos. Sci.*, 71, 1090-1104. http://dx.doi.org/10.1175/JAS-D-13-0182.1.

Examples

```
# generate a box function with white noise
set.seed(123)
x1 = cbfs(type = "box", sigma = 1)
# generate a box function with higher level noise
set.seed(123)
x2 = cbfs(type = "box", sigma = 3)
# plot them
par(mfrow=c(1,2))
plot(x1,type="l",xlab="t",ylab=expression(x[1]))
plot(x2,type="l",xlab="t",ylab=expression(x[2]))
```

`cbfs_red` *Generate an artificial event with red noise*

Description

This function generates a box, cliff-ramp, ramp-cliff or a sine function with red noise (AR(1)) as the background noise. Length of the generated event is 128.

Usage

```
cbfs_red(type = c("box", "rc", "cr", "sine"), A = 10, s = 1,
  coeff = 0.5)
```

Arguments

`type`	type of the event to be generated. There are four options: “box", “rc",“cr",“sine" representing a box, cliff-ramp, ramp-cliff or a sine function.
`A`	amplitude of the event; default is 10.
`s`	standard deviation of the AR(1) model innovations. Default is 1.
`coeff`	coefficient of the AR(1) process, which is used to control the level of red noise. Default is 0.5.

Value

an artificial event with red noise.

Examples

```
# generate a box function with red noise
set.seed(123)
x = cbfs_red(type = "box", coeff=0.5, s=1, A=10)
# plot it
plot(x,type="l",xlab="t",ylab="x")
```

`detrendc` *Conditionally detrend a time series*

Description

This function detrends a time series when its linear trend is more significant than a threshold.

Usage

```
detrendc(x, thres = 0.85)
```

Arguments

`x`	a vector or time series.
`thres`	a scalar specifying the threshold. When the adjusted R square coefficient of the linear fitting is larger than this threshold, the linear trend is substracted from the original time series. Default is 0.85.

Value

detrended `x`.

Examples

```
t=seq(0.001,1,0.001)
set.seed(123)
x=10*t+rnorm(1000)
dtrx=detrendc(x)
# plot the simulated x
plot(t,x,ty="l",xlab="t",ylab="x")
# plot the detrended x
lines(t,dtrx,col=2)
legend(0,12,legend=c("x","detrended x"),col=c(1,2),lty=1)
```

---

eventCluster *Cluster detected events*

Description

This function groups the detected events into clusters.

Usage

```
eventCluster(events, k0)
```

Arguments

events	an object of class 'events'.
k0	the number of clusters.

Details

The clustering is based on statistical characteristics of event. Each extracted event is first described using a feature vector, and then the events are clustered according to the Euclidean distances among the feature vectors. Note that before clustering, we apply principal component analysis (PCA) to the feature vectors to reduce the correlation as well as the dimension of the feature space.

Value

a list consisting of:

`cl`	a vector indicating which cluster each event belongs to.
`center`	a matrix which gives cluster centroids.
`pca`	PCA results for characteristics of the detected events.

References

Xiaozhe Wang, Kate Smith-Miles and Rob Hyndman (2005). Characteristic-Based Clustering for Time Series Data. *Data Mining and Knowledge Discovery*. 13(3), 335-364. http://dx.doi.org//10.1007/s10618-005-0039-x

Yanfei Kang, Danijel Belusic, Kate Smith-Miles (2014). Detecting and Classifying Events in Noisy Time Series. *J. Atmos. Sci.*, 71, 1090-1104. http://dx.doi.org/10.1175/JAS-D-13-0182.1.

Gregory S. Poulos, William Blumen, David C. Fritts, Julie K. Lundquist, Jielun Sun, Sean P. Burns, Carmen Nappo, Robert Banta, Rob Newsom, Joan Cuxart, Enric Terradellas, Ben Balsley, and Michael Jensen. CASES-99: A comprehensive investigation of the stable nocturnal boundary layer (2002). *Bulletin of the American Meteorological Society*, 83(4):555-581.

See Also

measures

Examples

```
##################################
#   An artificial example
##################################
set.seed(123)
n=128
types=c("box","rc","cr","sine")
shapes=matrix(NA,20,n)
for (i in 1:20){
  shapes[i,]=cbfs(type=types[sample(1:4,1)])
}
whitenoise=ts2mat(rnorm(128*20),128)
# generate x which randomly combine the four types of events with each
# two of them separated by noise
x=c(rnorm(128),t(cbind(shapes,whitenoise)))
# plot(x,ty="l")
# specify a sliding window size
w=128
# specify a significant level
alpha=0.05
# event detection
# events=eventDetection(x,w,"white",parallel=FALSE,alpha, "art")
# clustering
# cc=eventCluster(events,4)
# myclkm=cc$cl
##################################
```

```
#   CASES-99 dataset (9.5m)
###################################
# a sliding window length chosen by the user
w=120;
# specify a significant level
alpha=0.05
data(CASES99)
# CASESevents=eventDetection(CASES99,w,"red",FALSE,0.05,"real")
# cc=eventCluster(CASESevents,3)
# cc$center
# myclkm=cc$cl
# visualise the clustering in 2-dimension PCA space
# pc.cr=cc$pca
# pca.dim1 <- pc.cr$scores[,1]
# pca.dim2 <- pc.cr$scores[,2]
# plot(pca.dim1,pca.dim2,col=myclkm+1,
# + main="PCA plots for k-means clustering",pch=16)
```

`eventDetection` *Detect events from time series*

Description

This function finds events from a time series.

Usage

```
eventDetection(x, w, noiseType = c("white", "red"),
   parallel = FALSE, alpha = 0.05, data = c("art", "real"))
```

Arguments

`x`	a vector or time series.
`w`	size of the sliding window.
`noiseType`	background noise type assumed for x. There are two options: white noise or red noise.
`parallel`	logical, if TRUE then codes are executed in parallel using **foreach** package. The user must register a parallel backend to use by the **doMC** package.
`alpha`	the significance level. When the noise test p value of the subsequence is smaller than this significance level, it is defined as a potential event. Default is 0.05.
`data`	type of data being analysed. There are two options: 'art' if analysed data is artificial data and 'real' if analysed data is real world turbulence data. Please see the details in Kang et al. (2014).

Value

an object of class 'events' with the components listed below:

`x`	the original time series.
`start`	a vector consisting of starting points of events.
`end`	a vector consisting of ending points of events.
`nevents`	number of detected events.

References

Yanfei Kang, Danijel Belusic, Kate Smith-Miles (2014): Detecting and Classifying Events in Noisy Time Series. *J. Atmos. Sci.*, 71, 1090-1104. http://dx.doi.org/10.1175/JAS-D-13-0182.1.

Gregory S. Poulos, William Blumen, David C. Fritts, Julie K. Lundquist, Jielun Sun, Sean P. Burns, Carmen Nappo, Robert Banta, Rob Newsom, Joan Cuxart, Enric Terradellas, Ben Balsley, and Michael Jensen. CASES-99: A comprehensive investigation of the stable nocturnal boundary layer (2002). *Bulletin of the American Meteorological Society*, 83(4):555-581.

See Also

`noiseTests`, `eventExtraction`, `plotevents`

Examples

```
###################################
#   1st art eg (white noise)
###################################
set.seed(123)
n=128
types=c("box","rc","cr","sine")
shapes=matrix(NA,20,n)
for (i in 1:20){
  shapes[i,]=cbfs(type=types[sample(1:4,1)])
}
whitenoise=ts2mat(rnorm(128*20),128)
# generate x which randomly combine the four types of events with each
# two of them separated by noise
x=c(rnorm(128),t(cbind(shapes,whitenoise)))
```

```
plot(x,ty="l")
# specify a sliding window size and significant level
# w=128; alpha=0.05
# events=eventDetection(x,w,"white",parallel=FALSE,alpha,"art")
#################################
#   2nd art eg (red noise)
#################################
set.seed(123)
# set a red noise level
coeff=0.5;s=1
# generated x with red noise as the background; this time series is the
# one used in Kang et al. (2014)
x=c(arima.sim(list(order = c(1,0,0),ar=coeff),n=500,sd=s),
+  cbfs_red("rc"),
+  arima.sim(list(order = c(1,0,0),ar=coeff),n=400,sd=s),
+  cbfs_red("cr"),
+  arima.sim(list(order = c(1,0,0),ar=coeff),n=400,sd=s),
+  cbfs_red("box"),
+  arima.sim(list(order = c(1,0,0),ar=coeff),n=400,sd=s),
+  cbfs_red("sine"),
+  arima.sim(list(order = c(1,0,0),ar=coeff),n=1000,sd=s),
+  arima.sim(list(order = c(1,0,0),ar=0.8),n=1100,sd=4))
# specify a sliding window size and significant level
# w=128; alpha=0.05
# event detection
# events=eventDetection(x,w,"red",parallel=FALSE,alpha,"art")
#################################
#   CASES-99 dataset (9.5m)
#################################
# window size which needs to be chosen by the user
w=120
```

```
# specify a significant level
alpha=0.05
# event detection from CASES99 data
# data(CASES99)
# CASESevents=eventDetection(CASES99,w,"red",FALSE,alpha,"real")
```

`eventExtraction` *Extract events from time series*

Description

This function returns the starting and ending points of events according to the noise test results from a time series.

Usage

```
eventExtraction(tests, w, alpha = 0.05)
```

Arguments

`tests`	test p values from the noist tests for the subsequences.
`w`	sliding window size.
`alpha`	the significance level. When the noise test p value of the subsequence is smaller than this significance level, it is a potential event. Default is 0.05.

Value

a list consisting:

`start`	a vector consisting of starting points of events.
`end`	a vector consisting of ending points of events.
`tests`	smoothed test p value series.
`nevents`	number of detected events.

References

Yanfei Kang, Danijel Belusic, Kate Smith-Miles (2014): Detecting and Classifying Events in Noisy Time Series. *J. Atmos. Sci.*, 71, 1090-1104. `http://dx.doi.org/10.1175/JAS-D-13-0182.1`.

`measures` *Calculate statistical characteristics of an event*

Description

This function calculates statistical characteristics for detected events.

Usage

```
measures(x)
```

Arguments

`x`	a time series

Details

Measures used here are standard deviation, kurtosis, skewness, HD (the absolute Difference between averages of the first and second Half), nonsmoothness,

test statistic of PP test and ZA test, and maximum, minimum, and kurtosis of the first-order difference of the events. Please see the reference for details (Kang et al. 2014).

Value

a vector consisting of statistical characteristics of event `x`

References

Yanfei Kang, Danijel Belusic, Kate Smith-Miles (2014). Classes of structures in the stable at- mospheric boundary layer. Submitted to Quarterly Journal of the Royal Meteorological Society.

See Also

`eventCluster`

Examples

```
set.seed(123)
n=128
measures(cbfs("box"))
measures(cbfs("sine"))
```

`noiseTests` *Perform noise tests for a time series*

Description

This function performs noise tests on the sliding subsequences extracted from a time series.

Usage

```
noiseTests(x, w, noiseType = c("white", "red"), parallel = FALSE)
```

Arguments

`x`	a vector or a time series.
`w`	a scalar specifying the size of the sliding window.
`noiseType`	background noise assumed for `x`. There are two options: “white" or “red" .
`parallel`	logical, if TRUE then codes are executed in parallel using the **foreach** package. The user must register a parallel backend to use by the **doMC** package.

Details

When using this function, the user needs to choose the background noise type via `noiseType` according to the application context. In atmospheric turbulence, red noise is used. We first use the Phillips-Perron (PP) Unit Root Test to test for the unit root process. For the stationary processes, red noise tests are performed to test for events. For those cases tested to be unit root processes, we have to take into consideration a special situation when there is a structural break in the process. The reason comes from the difficulty for PP test to distinguish random walk processes from a stationary process contaminated by a structural break, both of which result in non-rejection of

the null hypothesis. Random-walk processes are not considered as events since they are known to be brownian noise, but stationary processes with structure breaks are, so it is essential to distinguish them. To this end, an additional test called Zivot & Andrews (ZA) unit root test is introduced.This test allows for a structural break in either the intercept or in the slope of the trend function of the underlying series. Rejection of the null hypothesis indicates a potential event (stationary process with a structural break). Random walk processes result in non-rejection of the null hypothesis.

Value

test p value series for the time series x.

References

Pierre Perron (1998). Trends and random walks in macroeconomic time series: Further evidence from a new approach. *Journal of economic dynamics and control*, 12(2), 297-332. http://dx.doi.org/10.1016/0304-3932(82)90012-5.

Eric Zivot and Donald W K Andrews (1992). Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of Business & Economic Statistics*, 20(1), 25-44. http://dx.doi.org/10.1198/073500102753410372.

Yanfei Kang, Danijel Belusic and Kate Smith-Miles (2014). Detecting and Classifying Events in Noisy Time Series. *J. Atmos. Sci.*, 71, 1090-1104. http://dx.doi.org/10.1175/JAS-D-13-0182.1.

See Also

eventExtraction, plotevents

Examples

```
set.seed(123)
n=128
types=c("box","rc","cr","sine")
shapes=matrix(NA,20,n)
for (i in 1:20){
  shapes[i,]=cbfs(type=types[sample(1:4,1)])
}
whitenoise=ts2mat(rnorm(128*20),128)
# generate x which randomly combine the four types of events with each
# two of them separated by noise
x=c(t(cbind(shapes,whitenoise)))
plot(x,ty="l")
w=128
# execute loops sequentially
tests=noiseTests(x,w,"white",parallel=FALSE)
# execute loops in parallel using doMC package (for non-Windows users)
# tests=noiseTests(x,w,"white",parallel=TRUE)
```

---

plotevents *Plot the detected events*

Description

This function plots the detected events from a time series.

Usage

```
plotevents(events, cluster = FALSE, mycl, ...)
```

Arguments

`events`	an object of class 'events'.
`cluster`	logical, if TRUE then the detected events are highlighted using different colors for different clusters
`mycl`	a vector specifying which cluster each event belongs to
`...`	other arguments that can be passed to plot

Value

...

References

Yanfei Kang, Danijel Belusic and Kate Smith-Miles (2014). Detecting and Classifying Events in Noisy Time Series. *J. Atmos. Sci.*, 71, 1090-1104. `http://dx.doi.org/10.1175/JAS-D-13-0182.1`.

See Also

`noiseTests`, `eventExtraction`, `eventDetection`

Examples

```
##################################
#   1st art eg (white noise)
##################################
set.seed(123)
n=128
types=c("box","rc","cr","sine")
shapes=matrix(NA,20,n)
```

```
for (i in 1:20){
  shapes[i,]=cbfs(type=types[sample(1:4,1)])
}
whitenoise=ts2mat(rnorm(128*20),128)
# generate x which randomly combine the four types of events with each
# two of them separated by noise
x=c(rnorm(128),t(cbind(shapes,whitenoise)))
plot(x,ty="l")
w=128; alpha=0.05
# event detection
# events=eventDetection(x,w,"white",FALSE,alpha,"art")
# clustering events
# cc=eventCluster(events,4)
# myclkm=cc$cl
# plot the clustered events
# plotevents(events,cluster=TRUE, myclkm)
###################################
#   2nd art eg (red noise)
###################################
set.seed(123)
# generate a time series with red noise; this is the same with the one
# used  in Kang et al. (2014)
coeff=0.5;s=1
x=c(arima.sim(list(order = c(1,0,0),ar=coeff),n=500,sd=s),
+  cbfs_red("rc"),
+  arima.sim(list(order = c(1,0,0),ar=coeff),n=400,sd=s),
+  cbfs_red("cr"),
+  arima.sim(list(order = c(1,0,0),ar=coeff),n=400,sd=s),
+  cbfs_red("box"),
+  arima.sim(list(order = c(1,0,0),ar=coeff),n=400,sd=s),
+  cbfs_red("sine"),
```

```
+  arima.sim(list(order = c(1,0,0),ar=coeff),n=1000,sd=s),
+  arima.sim(list(order = c(1,0,0),ar=0.8),n=1100,sd=4))
w=128; alpha=0.05
# event detection
# events=eventDetection(x,w,"red",parallel=FALSE,alpha,"art")
# plot events without clustering
# plotevents(events)
```

`ts2mat` *Reshape a vector into a matrix*

Description

This function reshapes a vector into a matrix whose row elements are taken from the vector. Orders of elements keep unchanged from the vector.

Usage

```
ts2mat(x, w)
```

Arguments

`x`	a vector or a time series
`w`	a number specifying number of columns of the matrix

Value

a matrix

Examples

```
x=ts2mat(c(1:(128*20)),128)
dim(x)
x[1,1:20]
```

`ur.za.fast` *Unit root test for events considering a structrual break*

Description

This function performs the Zivot & Andrews unit root test, which allows a break at an unknown point in either the intercept, the linear trend or in both.

Usage

```
ur.za.fast(y, model = c("intercept", "trend", "both"),
     lag = NULL)
```

Arguments

`y`	a vector or a time series.
`model`	Three choices: “intercept", “trend" or “both".
`lag`	a scalar chosen as lag.

Details

This function is written refering to the `ur.za` function in the **urza** package (Pfaff 2008), but it speeds up executation using the **RcppArmadillo** package. Allowing a structrual break, this function returns flag to be 0 if the time series is stationary and 1 if it is a unit root process.

Value

a list consisting of:

`flag`	0 if the time series is is stationary; 1 if it is a unit root process.
`teststat`	ZA unit root test statistic.

References

Eric Zivot and Donald W K Andrews (1992). Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of Business & Economic Statistics*, 20(1), 25-44. http://dx.doi.org/10.1198/073500102753410372.

Pfaff, Bernhard (2008). Analysis of Integrated and Cointegrated Time Series with R. Second Edition. Springer, New York. http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-75966-1.

See Also

`noiseTests`

Examples

```
# this is a box function
set.seed(123)
x=cbfs_red("box")
ur.za.fast(x,"both")
# this is a cliff-ramp
set.seed(123)
x=cbfs_red("cr")
ur.za.fast(x,"both")
```

```
# this is a random walk process
set.seed(123)
x=cumsum(rnorm(300))
ur.za.fast(x,"both")
```