Detecting and Classifying Events in Noisy Time Series

YANFEI KANG, DANIJEL BELUŠIĆ, AND KATE SMITH-MILES

School of Mathematical Sciences, Monash University, Melbourne, Victoria, Australia

(Manuscript received 19 June 2013, in final form 22 September 2013)

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.

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 are 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 wavenumber. This is the characteristic shared with red noise, and as a result, atmospheric time series are frequently modeled using a first-order autoregressive [AR(1)] process. 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, 2007; 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 then to 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 2011). Such situations do not allow for the above-mentioned approach

Corresponding author address: Danijel Belušić, School of Mathematical Sciences, Building 28, Monash University, Clayton, VIC 3800, Australia.
E-mail: danijel.belusic@monash.edu

DOI: 10.1175/JAS-D-13-0182.1

but require 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 that a significant number of structures could be explained or characterized by such an 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 preassumed or predefined 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 2. The method is first tested on artificially generated time series with different complexities of noise—white and red noise—hence progressing toward the real-world atmospheric conditions. This is detailed in section 3. The method is then applied to atmospheric turbulence data, as discussed in section 4. The basic assumptions of the method are further tested in section 5, and the conclusions are drawn in section 6.

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 and Lin 2005). Therefore, if the goal of analysis is to extract and classify events from time series, then 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.

a. 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.

1) 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

$$Q = n(n+2)\sum_{k=1}^{h}\frac{(\widehat{\rho_k}^2)}{n-k},$$

where n is the sample size, $\widehat{\rho_k}$ is the sample autocorrelation 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. 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.

2) RED NOISE TEST

Red noise is modeled as a first-order autoregressive process. Given a time series with red noise, the white noise test from section 2a(1) would not recognize any part of the time series as noise; hence, a separate test for red noise needs to be introduced.

Equation (1) defines an AR(1) process as a first-order autocorrelation model with the error term represented by a white noise process:

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

where $x(t)$ is a time series, ϕ is the first-order autocorrelation coefficient ($0 < \phi < 1$), and ϵ (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 (von 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 nonstationary, a stationarity test is applied first to the given time series $x(t)$. A nonparametric 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., it is nonstationary) against a stationary alternative. The test has been implemented in many statistical software packages, such as R package statistics (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 nonstationary 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 2a(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. First, 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., von 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., von Storch and Zwiers 1999). Defining events in the present method as non-AR(1) processes means that the events are nonstationary, oscillatory, and/or nonlinear 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.

b. 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

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

for $1 \le q \le m - l + 1$, where l is the sliding window size, which is also the length of the subsequence. The sliding window sizes l are prechosen 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 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: $p_1, p_2, \ldots, p_{m-l+1}$. When the subsequence test p value is smaller than a predefined significance level, we reject H_0 from section

2a(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 Δ_t, a noise test on sliding subsequences will in general return consecutive potential events from the time point $t_0 - \Delta_{t_1}$ to $t_0 + \Delta_{t_2}$, where $\Delta_{t_{1,2}} \leq \Delta_t$. 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 that 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. Here, we use $\alpha = 0.05$. Assume there exists a consecutive progression of p values $p_s, p_{s+1}, \ldots, p_e$, which satisfies

(i) $p_i \leq \alpha, i = s, s+1, \ldots, e$ and
(ii) $e - s \geq l/2$.

Then, we define the middle subsequence $x_{\lfloor (e+s)/2 \rfloor}(t)$ as ''the event'' for which we are searching, 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 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 1993) 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 is 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 1993). 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 3b(3).

c. *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 that are easier for the analysts 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. 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, nonlinearity (Wang et al. 2006), serial correlation (Wang et al. 2006), trend, period, kurtosis, skewness, and nonsmoothness, as well as the maximum, minimum, standard deviation, serial correlation, and kurtosis of the first-order difference of the subsequences. The

period for $x(t)$ 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 that is larger than the critical value $1.96/\sqrt{n}$ (Enders 2003).
- If no peak satisfies the condition above, then there is no periodicity in $x(t)$.

The nonsmoothness 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 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.

d. 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, then 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 5b.

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 are inserted in noise of various levels and characteristics. The complexity of noise increases toward red noise, which is a step that leads toward applications to real-world datasets.

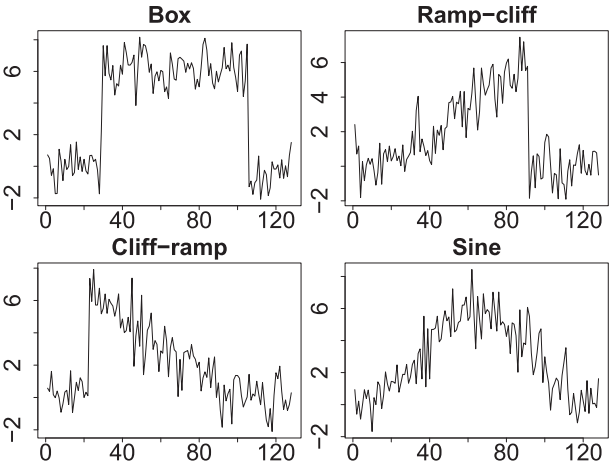

FIG. 1. Examples of box, ramp–cliff, cliff–ramp, and sine shapes.

a. 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. To make the task of finding shapes more challenging and hence closer to realistic data, the shapes have variable lengths smaller than 128 points. The start and end points of shapes vary: a from 16 to 32 points and b from 64 to 128 points. Figure 1 shows an instance of each of the four shapes with some Gaussian noise added.

1) 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 points). 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 standard deviation $\sigma = 1$ is shown in Fig. 2 (top).

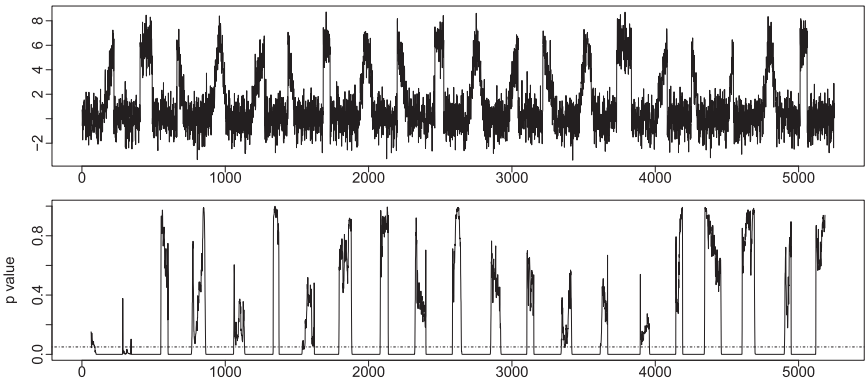

FIG. 2. Artificial time series with (top) background white noise with $\sigma = 1$ and (bottom) the corresponding Ljung–Box test p values. The dotted–dashed line represents the threshold $\alpha = 0.05$. A p value smaller than α indicates a possible shape. Notice that a single p value corresponds to a subsequence of length $l = 128$ points, and the location of p in the time series corresponds to the central point of the subsequence.

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 3σ. [See Fig. 5 (top) for the artificial time series with noise level $N(0, 3^2)$.]

3) TIME SERIES WITH RED NOISE

As a step toward 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. 6.)

b. Results

1) BACKGROUND WHITE NOISE

In this case, we know that the length of the embedded shape regions is 128 points, so we use a sliding window with the same length ($l = 128$ points) 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, nonlinearity, serial correlation, and trend, and maximum, minimum, standard deviation, and serial correlation of the first-order difference of the subsequences. Figure 2 (bottom) depicts the p-value series of the sliding subsequences of length 128 points extracted from the artificially generated time series in Fig. 2 (top). 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. 3. As can be seen from the 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. 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. 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.

2) 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

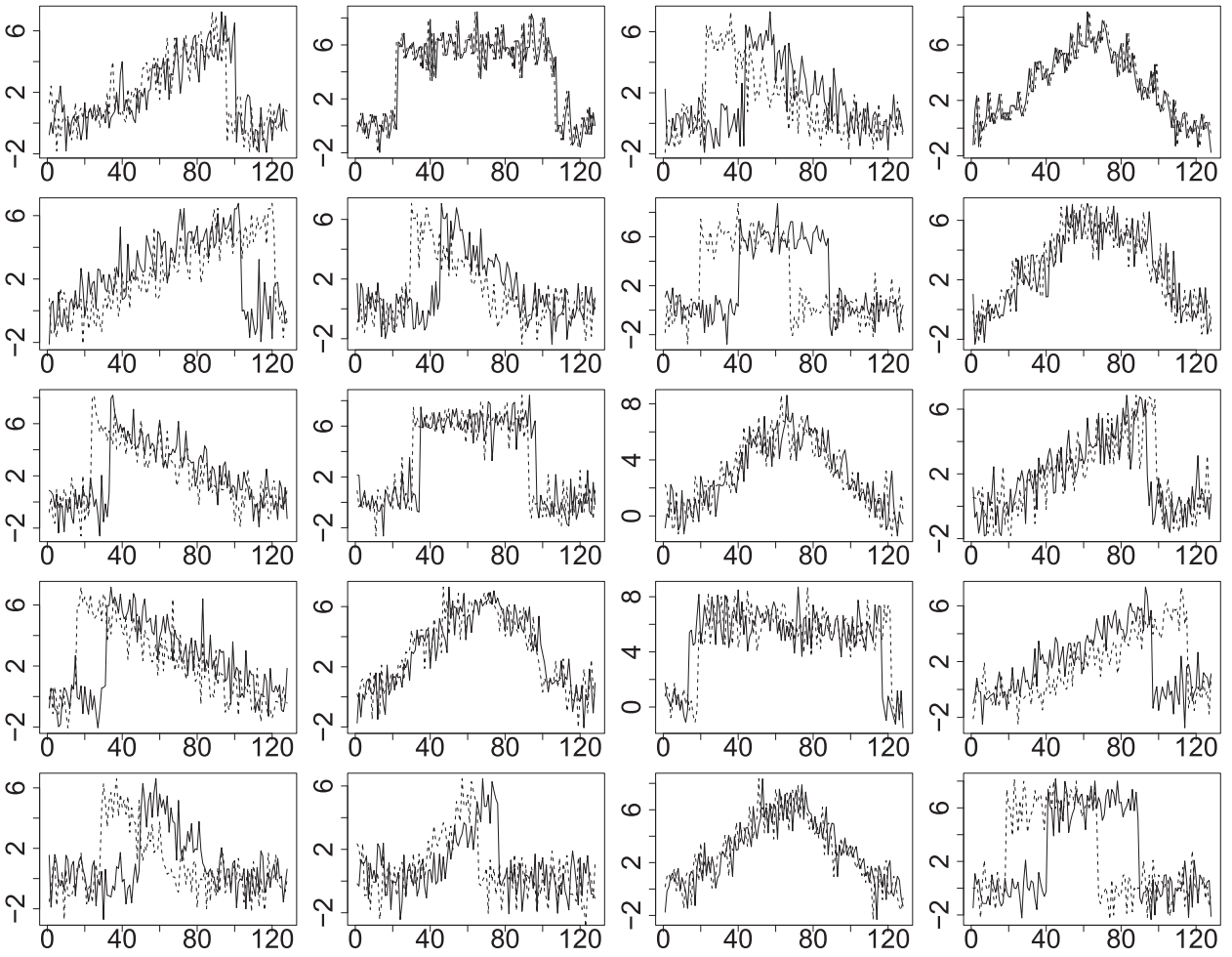

FIG. 3. The 20 shapes extracted from the artificial time series shown in Fig. 2 (top). The dashed lines in the background show the original shapes used to generate the time series.

challenging with higher noise values because of the difficulties in distinguishing shapes from noise. To illustrate the results, Fig. 5 (bottom) shows the p value series corresponding to the time series with 3σ 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 3 times 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.

3) BACKGROUND RED NOISE

The red noise test is applied to the artificial data with four shapes and the background red noise, shown in Fig. 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. 6, top). Figure 6 (bottom) 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

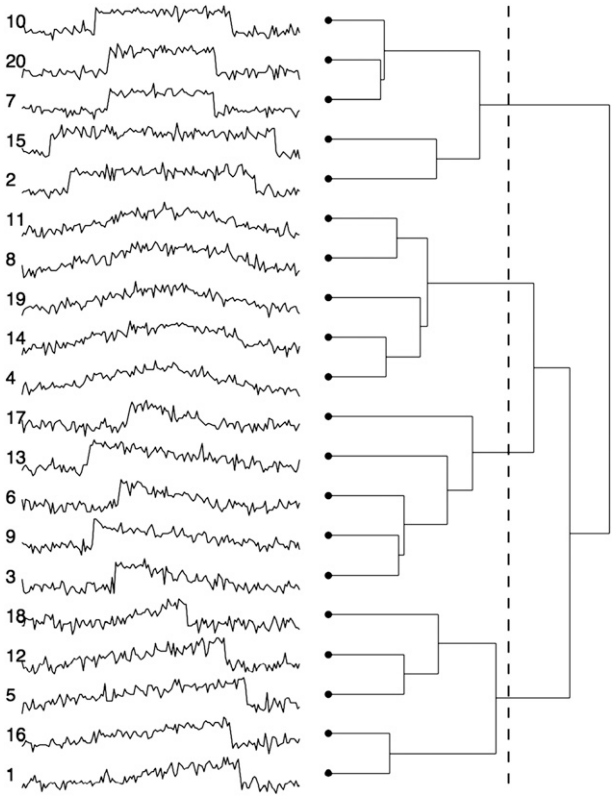

FIG. 4. Dendrogram from hierarchical clustering of the extracted shapes based on features; the vertical line shows where the binary tree is cut to get the four basic types of shapes.

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.

4. Application to real-world turbulence data

a. Data description

Data from the 1999 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, 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, 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).

We use 1-s averages of the thermocouple and sonic anemometer data. The thermocouple at the seventh level (9.5 m) from 1100 LST 5 October to 1100 LST 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.

b. 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}, ..., p_e of the corresponding subsequences, which satisfy the two rules in the definition of the event (see section 2b), the event would be chosen as the middle subsequence for simple artificial data with known lengths of shapes In the real-world

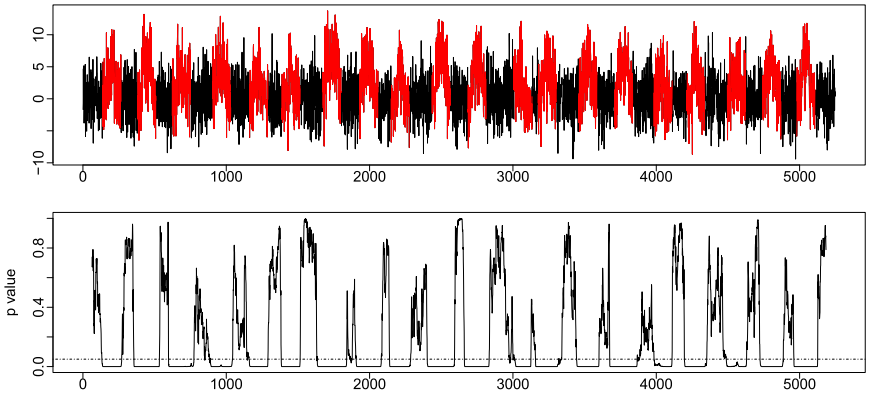

FIG. 5. As in Fig. 2, except that the white noise level is increased to 3 times as before. (top) The detected shapes are colored red.

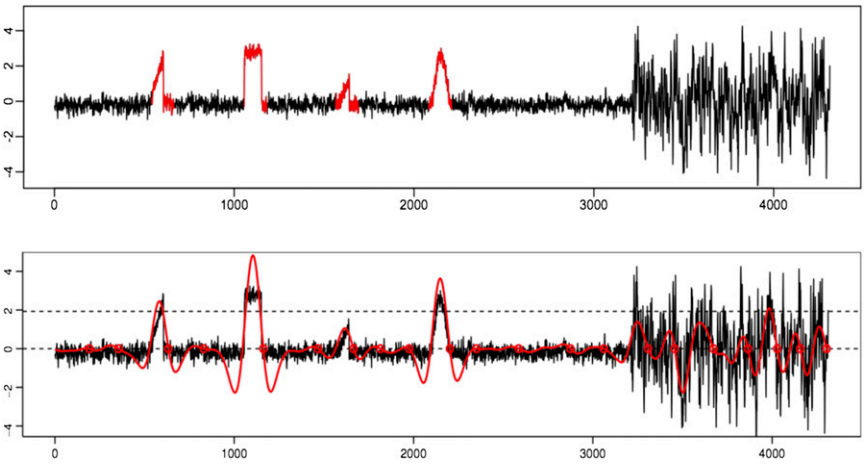

FIG. 6. Time series with background red noise and the comparison with wavelets. The first half of this artificial time series consists of four basic shapes and background red noise with the autocorrelation coefficient $\phi = 0.4$. The second half consists of two equal-length segments of pure red noise with two different values of ϕ: 0.4 and 0.8 and their $\epsilon(t)$ follows $N(0, 1)$ and $N(0, 4)$, respectively. (top) The color-coded parts show shapes detected using the present method. (bottom) 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 dashed line indicates 40% of the absolute maximum of the coefficients at this scale.

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 l are still significantly different than smaller-scale noise over the range l. So, to take into account a tradeoff 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 potential event. 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, l is the minimum length of a recognized event, and there is no upper limit to the length of an event.

Using $l = 120$ 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, and nonsmoothness, and maximum, minimum, and kurtosis of the first-order difference of the subsequences. Thus, the hierarchical clustering algorithm is supposed to cluster the 102 eight-dimensional feature vectors into groups to find similarities among them. However, correlation analysis on the 102 events shows that some of the eight features are correlated; for example, the correlation between the kurtosis and the kurtosis of the first-order difference is 0.91. Therefore, before clustering, we apply 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. 7. To make the groups clearly separated, the tree is cut into six clusters shown in the six sidebars in Fig. 7. The number of clusters was chosen subjectively by visualizing the heat map and examining the results for several different numbers of clusters.

c. *Characteristics of events*

The following demonstrates the advantages of clustering the events and illustrates that the underlying mechanisms are physically meaningful. Figure 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 $\mathrm{Ri} = (g/\theta_0)\partial\overline{\theta}/\partial z(\partial\overline{\mathbf{V}}/\partial z)^{-2}$, where g is the gravitational acceleration, θ is the potential temperature, $\mathbf{V}$ is the wind vector, and the overbar

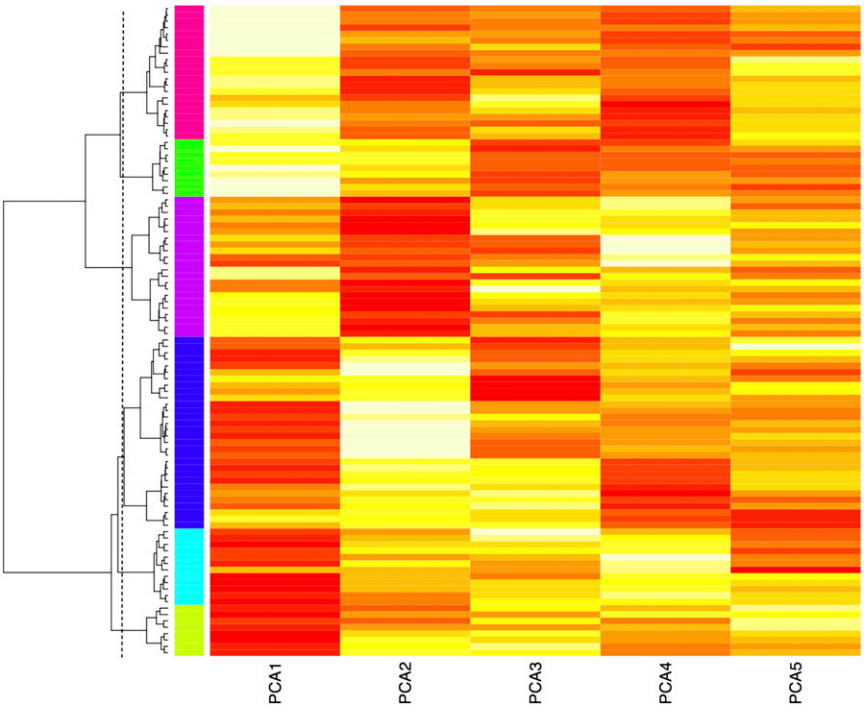

FIG. 7. Heat map for clustering of the extracted events. The hierarchical tree is cut into six clusters represented by the six sidebars. The vertical line shows where the binary tree is cut.

denotes the time average over the duration of an event. The vertical gradients are calculated using the 1.5- and 10-m levels for **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 indicates that clustering is able to group together structures with similar physical characteristics, given only a single-variable time series as the input.

Figure 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. 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 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.

To further visualize the clustering results, Fig. 10 depicts examples of events in each cluster, and Table 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 1). At the same time, cluster 5 contains periodic 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. Figure 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. Ramplike shapes in near-neutral conditions are

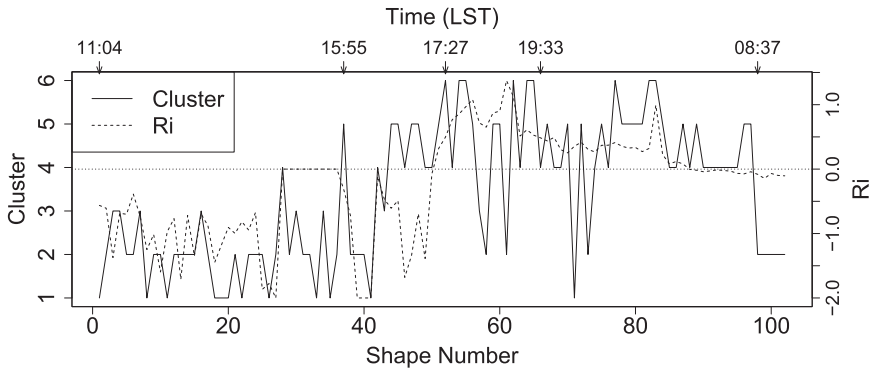

FIG. 8. Time evolution of the cluster number and Richardson number of extracted events. The horizontal dotted line denotes Ri = 0. The times on the top correspond to the event times when larger cluster transitions occur. The events were detected in the time series from 1100 LST 5 Oct to 1100 LST 6 Oct.

grouped in cluster 4. Meanwhile, the smoothest shapes go to cluster 6. The cluster 6 structures are deep and apparently wavelike.

A closer look at individual events further demonstrates their physical origin. For example, Fig. 11 shows the time–height cross sections of temperature and vertical velocity for a ramplike 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 the convectively unstable ABL studied by, for example, Wilczak (1984) and Williams and Hacker (1992). Another example is the event from cluster 6 that starts at 1917:06 LST 5 October. It is a part of the wavelike 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 2339:03 LST 5 October is again a part of the Kelvin–Helmholtz instability 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.

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 nonlinear 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 nonlinearity visually exhibit more expressed shapes. The second test investigates changes of event numbers after performing phase randomization (Dahlhaus et al. 2010)

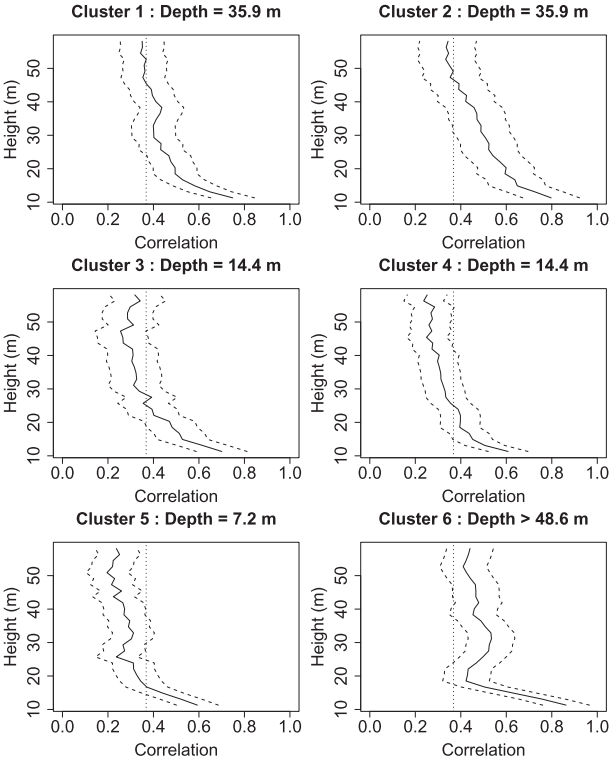

FIG. 9. Vertical correlations between the thermocouple at 9.5 m and those aloft, averaged over all events for each cluster. The dashed lines show a one-standard-deviation interval around the mean for each level. The vertical dotted lines represent e^{-1}. Titles show mean event depths for each cluster. When the depth is larger than the tower height, it is shown as >48.6 m; that is, larger than the difference between the highest thermocouple at 58.1 m and the reference thermocouple at 9.5 m.

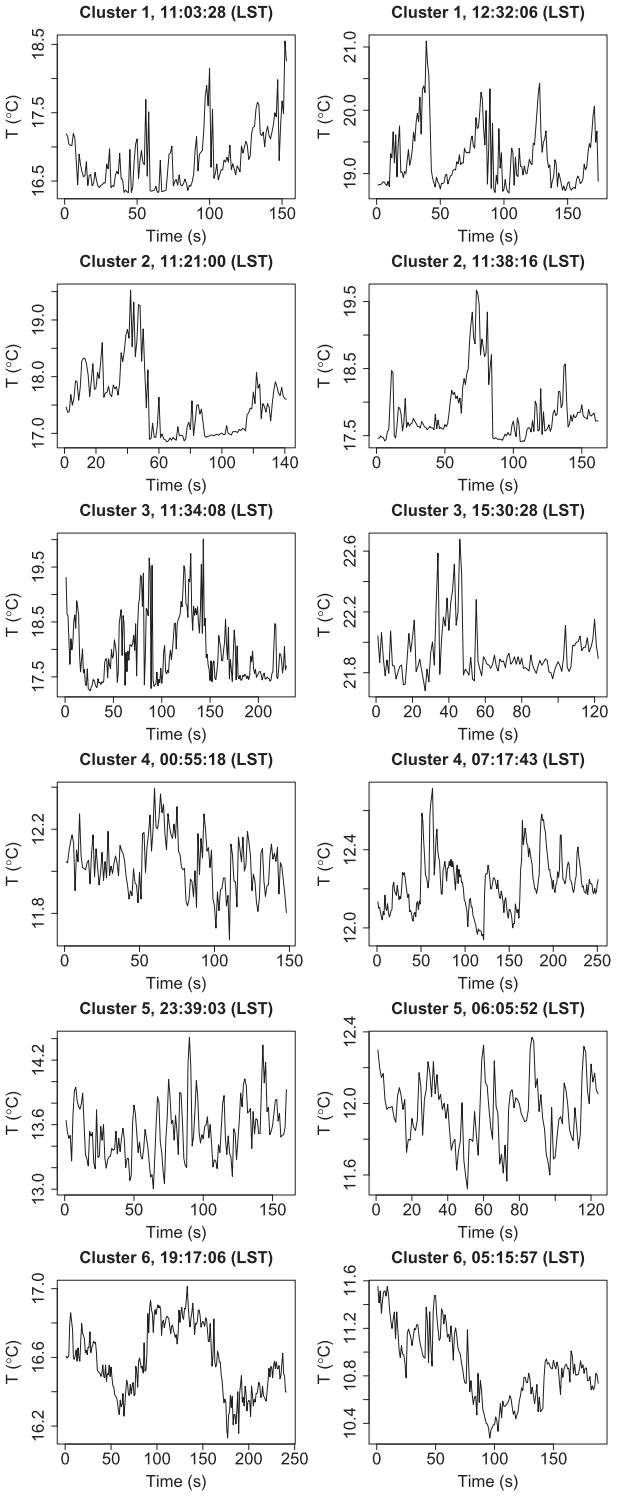

FIG. 10. Examples of events from the six clusters: two instances are shown from each cluster. The time of onset of an event is given in each title (the times are between 1100 LST 5 Oct and 1100 LST 6 Oct).

TABLE 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 nonsmoothness defined in section 2c—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

on a real-world time series. The two tests are presented in detail below.

a. *Artificial AR(1) time series with a nonlinear component*

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

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

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 nonlinear component into the 1000 generated time series (Gluhovsky and Agee 2007):

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

where a is a parameter that controls the nonlinearity of $y(t)$.

Figure 12 illustrates the changes that occur in the time series as the nonlinearity increases. It is clear even from simple visual inspection that individual shapes become more distinguishable with stronger nonlinearity. 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 nonlinearity. 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. Figure 13 shows that the percentage decreases with the increase of a. This means that time series become less AR(1)-like as the nonlinearity increases, which implies that the method correctly finds more events with stronger nonlinearity.

b. *Phase randomization*

As described in section 2d, phase randomization removes coherent structures from time series and can be

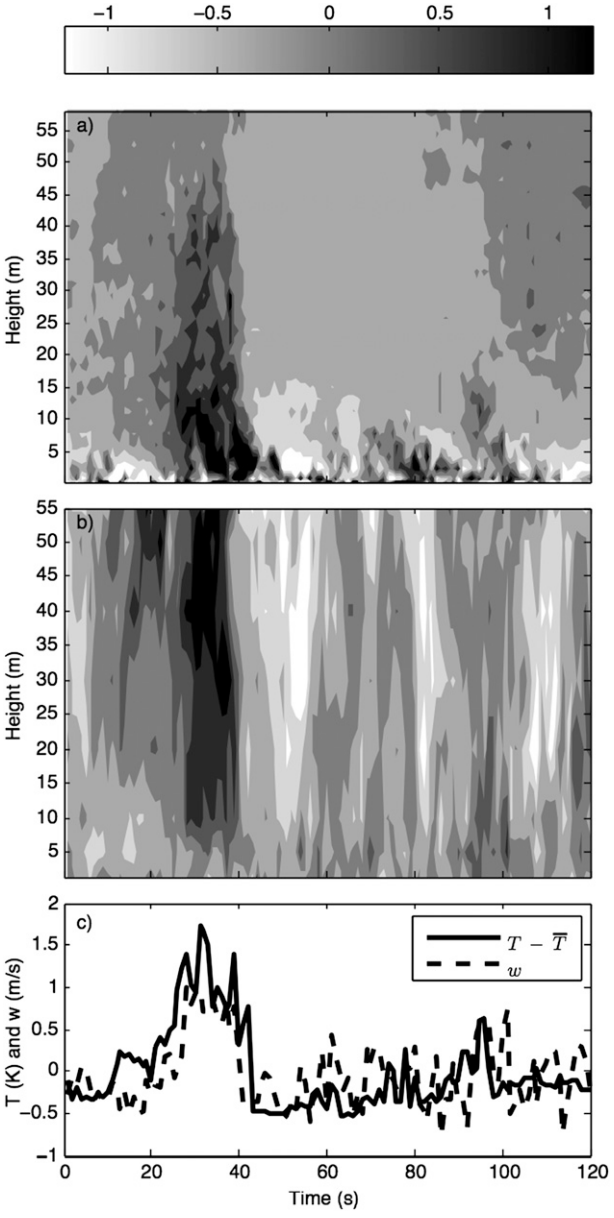

FIG. 11. Time–height cross section of the ramp structure that starts at 1138:16 LST 5 Oct showing (a) the temperature perturbation $[T(z,t) - \overline{T}(z)$, where the overbar denotes the time average over the event duration at each level] from the 34 thermocouples and (b) vertical velocity from the 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).

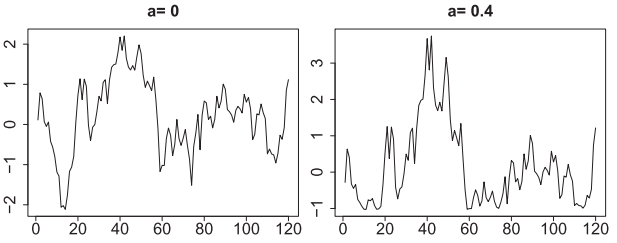

FIG. 12. AR(1) time series generated with different values of a.

used to validate the present method. The number of detected events is expected to be significantly smaller in the phase-randomized data compared with the original data. It should be noted that the present method 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. 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 ramplike coherent structures dominate the flow (Wilczak 1984). The length of the chosen section of the time series is $N = 20\,000$. 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 $p_2(t)$. We obtain $p_1(t)$ from the unmodified data using the red noise test as in section 2b, 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.

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.

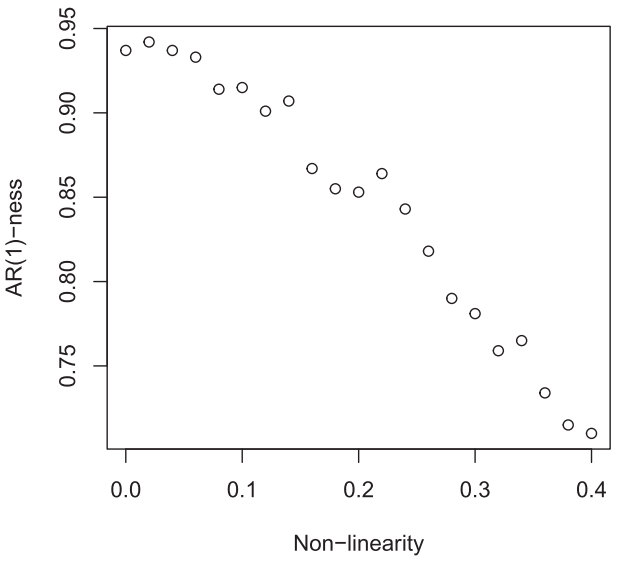

FIG. 13. Percentage of the time series characterized by $p > 0.05$ [i.e., those recognized as AR(1) or red noise] vs a.

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 nonnoise 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.

Acknowledgments. The valuable comments of Larry Mahrt and two anonymous reviewers are gratefully acknowledged. We thank Jielun Sun for providing the CASES-99 data and Eamonn Keogh for providing the MATLAB code that generates the Cylinder–Bell–Funnel datasets.

REFERENCES

Antonia, R. A., A. J. Chambers, C. A. Friehe, and C. W. V. Atta, 1979: Temperature ramps in the atmospheric surface layer. *J. Atmos. Sci.,* **36,** 99–108.

Baklanov, A. A., and Coauthors, 2011: The nature, theory, and modeling of atmospheric planetary boundary layers. *Bull. Amer. Meteor. Soc.,* **92,** 123–128.

Banerjee, A., J. J. Dolado, J. W. Galbraith, and D. Hendry, 1993: *Co-Integration, Error-Correction, and the Econometric Analysis of Non-Stationary Data.* Oxford University Press, 352 pp.

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

Belušić, D., and L. Mahrt, 2012: Is geometry more universal than physics in atmospheric boundary layer flow? *J. Geophys. Res.,* **117,** D09115, doi:10.1029/2011JD016987.

——, M. Hrastinski, Ž. Večenaj, and B. Grisogono, 2013: Wind regimes associated with a mountain gap at the northeastern Adriatic coast. *J. Appl. Meteor. Climatol.,* **52,** 2089–2105.

Blumen, W., R. Banta, S. P. Burns, D. C. Fritts, R. Newsom, G. S. Poulos, and J. Sun, 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. *Dyn. Atmos. Oceans,* **34,** 189–204, doi:10.1016/S0377-0265(01)00067-7.

Bolzan, M., F. Guarnieri, and P. C. Vieira, 2009: Comparisons between two wavelet functions in extracting coherent structures from solar wind time series. *Braz. J. Phys.,* **39,** 12–17.

Box, G. E. P., and D. A. Pierce, 1970: Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *J. Amer. Stat. Assoc.,* **65,** 1509–1526.

Campanharo, A. S. L. O., F. M. Ramos, E. E. N. Macau, R. R. Rosa, M. J. A. Bolzan, and L. D. A. Sá, 2008: Searching chaos and coherent structures in the atmospheric turbulence above the Amazon forest. *Philos. Trans. Roy. Soc,* **A366,** 579–589.

Chen, J., and F. Hu, 2003: Coherent structures detected in atmospheric boundary-layer turbulence using wavelet transforms at Huaihe River Basin, China. *Bound.-Layer Meteor.,* **107,** 429–444, doi:10.1023/A:1022162030155.

Chen, W., M. D. Novak, T. A. Black, and X. Lee, 1997: Coherent eddies and temperature structure functions for three contrasting surfaces. Part II: Renewal model for sensible heat flux. *Bound.-Layer Meteor.,* **84,** 125–147, doi:10.1023/a:1000342918158.

Chen, X., M. Wang, Y. Zhang, Y. Feng, Z. Wu, and N. E. Huang, 2013: Detecting signal from data with noise: Theory and applications. *J. Atmos. Sci.,* **70,** 1489–1504.

Chian, A. C.-L., R. A. Miranda, D. Koga, M. J. A. Bolzan, F. M. Ramos, and E. L. Rempel, 2008: Analysis of phase coherence in fully developed atmospheric turbulence: Amazon forest canopy. *Nonlinear Processes Geophys.,* **15,** 567–573.

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

Dahlhaus, R., J. Kurths, P. Maass, and J. Timmer, Eds., 2010: *Mathematical Methods in Time Series Analysis and Digital Image Processing.* 1st ed. Understanding Complex Systems, Springer, 308 pp.

Elsner, J. B., 2003: Tracking hurricanes. *Bull. Amer. Meteor. Soc.,* **84,** 353–356.

Enders, W., 2003: *Applied Econometric Times Series.* Wiley Series in Probability and Statistics, Vol. 804, Wiley, 480 pp.

Fu, T.-c., F.-l. Chung, V. Ng, and R. Luk, 2001: Pattern discovery from stock time series using self-organizing maps. *Proc. KDD 2001 Workshop on Temporal Data Mining,* San Francisco, CA, ACM, 26–29.

Ghil, M., and Coauthors, 2002: Advanced spectral methods for climatic time series. *Rev. Geophys.,* **40,** 1003, doi:10.1029/2000RG000092.

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

Gluhovsky, A., and E. Agee, 2007: On the analysis of atmospheric and climatic time series. *J. Appl. Meteor. Climatol.,* **46,** 1125–1129.

Holtslag, A. A. M., and Coauthors, 2013: Stable atmospheric boundary layers and diurnal cycles—Challenges for weather and climate models. *Bull. Amer. Meteor. Soc.,* 1691–1706.

Keogh, E., and S. Kasetty, 2002: On the need for time series data mining benchmarks: A survey and empirical demonstration. *Proc. Eighth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD),* Edmonton, Canada, ACM, 102–111.

——, and J. Lin, 2005: Clustering of time-series subsequences is meaningless: Implications for previous and future research. *Knowl. Inf. Syst.,* **8,** 154–177, doi:10.1007/s10115-004-0172-7.

Kuznetsov, E., and V. Zakharov, 2000: Nonlinear coherent phenomena in continuous media. *Nonlinear Science at the Dawn of the 21st Century,* P. L. Christiansen, M. P. Sorensen, and A. C. Scott, Eds., Lecture Notes in Physics, Vol. 542, Springer-Verlag, Berlin.

Mahrt, L., 2011: Surface wind direction variability. *J. Appl. Meteor. Climatol.,* **50,** 144–152.

Nappo, C., J. Sun, L. Mahrt, and D. Belušić, 2014: Determining wave–turbulence interactions in the stable boundary layer. *Bull. Amer. Meteor. Soc.,* in press.

Perron, P., 1988: Trends and random walks in macroeconomic time series: Further evidence from a new approach. *J. Econ. Dyn. Control,* **12,** 297–332.

Pope, M., C. Jakob, and M. J. Reeder, 2009: Objective classification of tropical mesoscale convective systems. *J. Climate,* **22,** 5797–5808.

Poulos, G. S., and Coauthors, 2002: CASES-99: A comprehensive investigation of the stable nocturnal boundary layer. *Bull. Amer. Meteor. Soc.,* **83,** 555–581.

Provenzale, A., L. Smith, R. Vio, and G. Murante, 1992: Distinguishing between low-dimensional dynamics and randomness in measured time series. *Physica D,* **58** (1–4), 31–49, doi:10.1016/0167-2789(92)90100-2.

R Core Team, 2013: *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, 3551 pp.

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

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

——, ——, ——, and ——, 2012b: Structure function analysis of two-scale scalar ramps. Part II: Ramp characteristics and surface renewal flux estimation. *Bound.-Layer Meteor.,* **145,** 27–44, doi:10.1007/s10546-012-9740-7.

Sun, J., L. Mahrt, R. M. Banta, and Y. L. Pichugina, 2012: Turbulence regimes and turbulence intermittency in the stable boundary layer during CASES-99. *J. Atmos. Sci.,* **69,** 338–351.

Thomas, C., and T. Foken, 2005: Detection of long-term coherent exchange over spruce forest using wavelet analysis. *Theor. Appl. Climatol.,* **80,** 91–104.

——, and ——, 2007: Organised motion in a tall spruce canopy: Temporal scales, structure spacing and terrain effects. *Bound.-Layer Meteor.,* **122,** 123–147.

——, J.-C. Mayer, F. X. Meixner, and T. Foken, 2006: Analysis of low-frequency turbulence above tall vegetation using a Doppler sodar. *Bound.-Layer Meteor.,* **119,** 563–587.

Tsay, R. S., 2005: *Analysis of Financial Time Series.* 2nd ed. Wiley-Interscience, 576 pp.

von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research.* Cambridge University Press, 496 pp.

Wang, X., K. A. Smith, and R. J. Hyndman, 2006: Characteristic-based clustering for time series data. *Data Min. Knowl. Discovery,* **13,** 335–364, doi:10.1007/s10618-005-0039-x.

Weber, R. O., and P. Kaufmann, 1995: Automated classification scheme for wind fields. *J. Appl. Meteor.,* **34,** 1133–1141.

Wilczak, J. M., 1984: Large-scale eddies in the unstably stratified atmospheric surface layer. Part I: Velocity and temperature structure. *J. Atmos. Sci.,* **41,** 3537–3550.

Williams, A., and J. Hacker, 1992: The composite shape and structure of coherent eddies in the convective boundary layer. *Bound.-Layer Meteor.,* **61,** 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, 394 pp.