Despiking Acoustic Doppler Velocimeter Data

Derek G. Goring[1] and Vladimir I. Nikora[2]

Abstract: A new method for detecting spikes in acoustic Doppler velocimeter data sequences is suggested. The method combines three concepts: (1) that differentiation enhances the high frequency portion of a signal, (2) that the expected maximum of a random series is given by the Universal threshold, and (3) that good data cluster in a dense cloud in phase space or Poincaré maps. These concepts are used to construct an ellipsoid in three-dimensional phase space, then points lying outside the ellipsoid are designated as spikes. The new method is shown to have superior performance to various other methods and it has the added advantage that it requires no parameters. Several methods for replacing sequences of spurious data are presented. A polynomial fitted to good data on either side of the spike event, then interpolated across the event, is preferred by the authors.

DOI: 10.1061/(ASCE)0733-9429(2002)128:1(117)

CE Database keywords: Turbulence; Velocity; Measuring instruments; Data processing.

Introduction

Acoustic Doppler velocimeters (ADVs) have become our instrument of choice for measuring velocities in our outdoor ecohydraulics flume (Nikora et al. 1998) and in the field (e.g., Nikora and Goring 2000), where other measurement techniques such as laser Doppler anemometers are impractical. However, ADVs also have some disadvantages. One of them is the Doppler noise floor and we discussed ways that we combat this in Nikora and Goring (1998)—see also Voulgaris and Trowbridge (1998). Another problem with ADVs is spikes caused by aliasing of the Doppler signal—the phase shift between the outgoing and incoming pulse lies outside the range between $-180°$ and $+180°$ and there is ambiguity, causing a spike in the record. Such a situation can occur when the flow velocity exceeds the preset velocity range or when there is contamination from previous pulses reflected from the boundaries of complex geometries (e.g., cobbles on the bed of a stream). Unfortunately, some of these spikes look remarkably similar to natural fluctuations in the velocity (Fig. 1).

In this article, we consider a number of different ways to detect spikes and how to deal with them. For the cases of a single-point spike, relatively simple despiking algorithms have proved satisfactory, but the situation with multipoint spikes, as illustrated in Fig. 1, has proved to be much more difficult. We have visited and revisited the problem several times, each time developing a solution, only to find that it does not work with the next set of ADV data. The method we have finally developed that works successfully on all of our ADV data is an amalgam of several ideas: that differentiating a signal enhances the high-frequency components (e.g., Roy et al. 1999); that the maximum of a white noise sequence is given by the Universal threshold (Donoho and Johnstone 1994; Katul and Vodakovic 1998); and the use of Poincaré maps (e.g., Abarbanel 1995; Addison 1997). The method uses the principle that good data lie within a cluster and that any data point lying well outside that cluster must be suspected of being a spike. McKinney (1993) described a similar method and applied it to ocean wave records.

[1]Princ. Sci., National Institute of Water and Atmospheric Research, P.O. Box 8602, Christchurch, New Zealand.

[2]Princ. Sci., National Institute of Water and Atmospheric Research, P.O. Box 8602, Christchurch, New Zealand.

Note. Discussion open until June 1, 2002. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this technical note was submitted for review and possible publication on May 10, 2000; approved on June 28, 2001. This paper is part of the ***Journal of Hydraulic Engineering***, Vol. 128, No. 1, January 1, 2002. ©ASCE, ISSN 0733-9429/2002/1-117–126/$8.00+$.50 per page.

Methods

Despiking involves two steps: (1) detecting the spike and (2) replacing the spike. The two steps are independent so they are considered separately here. Indeed, in most cases, the methods for spike detection described in the next section can be mixed interchangeably with any of the methods for spike replacement in the following section. However, for the iterative methods, spike replacement can affect spike detection in the subsequent iterations.

Spike Detection

General

Electrical engineers are well aware of the problem of spurious data and have developed numerous methods for handling them. Otnes and Enochson (1978) cover the problem (which they call ''wild point editing'') in some detail. The first two methods described in this section come from this source (RC Filters Method and Tukey 53H Method described in detail below). The methods involve digital filtering to generate two time sequences, one of which is ''rough'' and the other ''smooth.'' For each point in the data sequence, if the difference between the rough and smooth data exceed a threshold, the point is deemed spurious.

The authors have developed an algorithm for despiking ADV records of turbulence velocities in streams (Acceleration Thresholding Method). The method has been used widely in our work (e.g., Nikora and Goring 2000). It is based on the postulate that under normal flow conditions the instantaneous acceleration in a stream must be of the same order or less than the acceleration of gravity g (otherwise sediment grains would be thrown about violently, which is contrary to observations). The algorithm calcu-

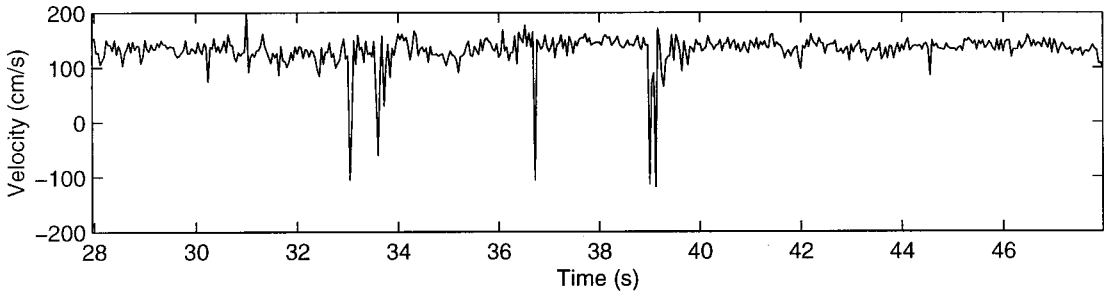

Fig. 1. Example of ADV data containing spikes from measurements near bed of stream

lates the accelerations from backward differences and identifies points as spikes if the acceleration exceeds one or two gravities. Surprisingly, it was found that this criterion was too severe and some apparently valid points were rejected as spikes, yet increasing the threshold allowed obvious spikes through. Therefore, an additional condition was introduced so that for a point to be a spike, the acceleration must exceed a threshold $\lambda_a g$ and the absolute deviation from the mean velocity of the point must exceed $k\sigma$, where λ_a is a relative acceleration threshold, σ is the standard deviation, and k is a factor, usually taken as 1.5. This method has proved very successful for records where the spikes are clearly different from fluctuations in the record, but for some records the choice of thresholds is very difficult and subjective. Thus, alternative methods have been investigated.

One of these (Wavelet Thresholding Method) arises from the landmark article by Donoho and Johnstone (1994) who introduced a new method for detecting and removing noise from a signal. Denoising is the converse of despiking, but the principle is similar. In their method the signal is transformed by orthogonal wavelet transformation, then the wavelet coefficients at the first scale, which contain most of the noise in the signal, are compared to a threshold. Those below the threshold are set to zero and those above the threshold are retained—this is called ''wavelet shrinkage.'' Thus, the inverse transform of the filtered wavelet coefficients is rendered noise free. Adapting this method to despiking simply entails rejecting the wavelet coefficients above the threshold, rather than those below the threshold as when denoising.

The threshold they apply arises from a theoretical result from normal probability distribution theory which says that for n independent, identically distributed, standard, normal, random variables ξ_i the expected absolute maximum is

$$E(|\xi_i|_{\max}) = \sqrt{2 \ln n} = \lambda_U \tag{1}$$

where λ_U is termed the Universal threshold. For a normal, random variable whose standard deviation is estimated by $\hat{\sigma}$ and the mean is zero, the expected absolute maximum is

$$\lambda_U \hat{\sigma} = \sqrt{2 \ln n}\hat{\sigma} \tag{2}$$

The failure of all of the above methods to adequately handle some ADV records prompted a radical re-evaluation of the strategy and the emergence of an idea to see how velocity and its derivatives look in phase space (we call this the Phase-Space Thresholding method), Fig. 2. It is immediately apparent that most of the data

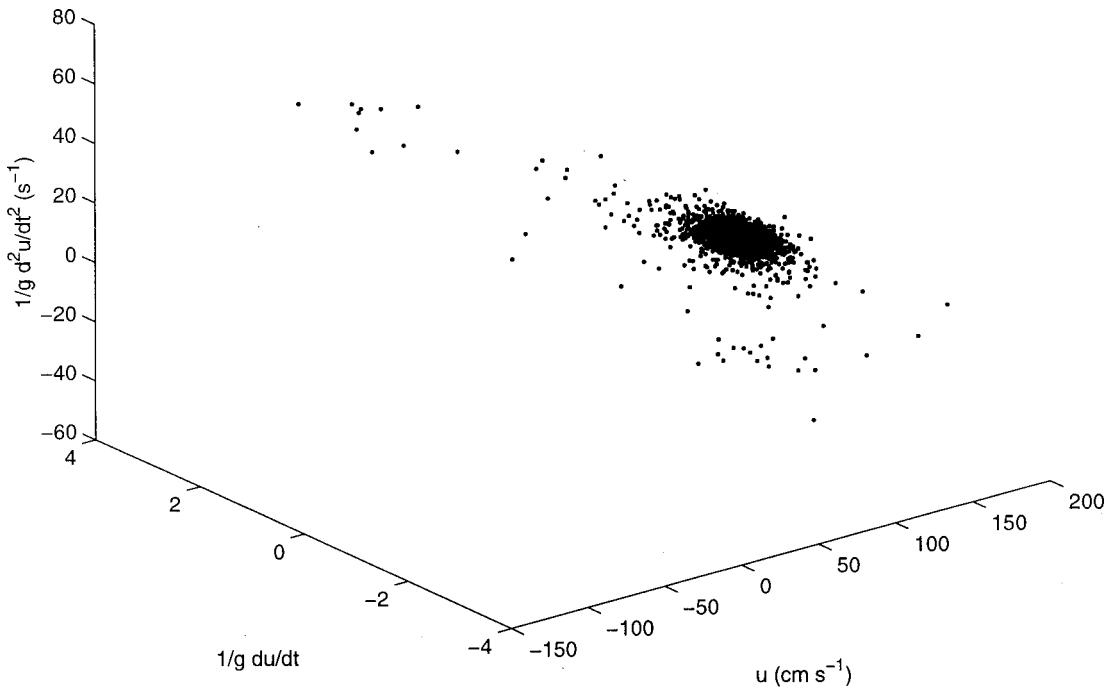

Fig. 2. Phase space showing cloud of data from ADV measurements, where derivatives have been scaled by acceleration of gravity

cluster in an ellipsoid cloud, and the spikes are separated from this cloud. The separation of the spikes is exaggerated for the derivatives because differentiation accentuates the high-frequency components (which the spikes belong to) compared to the low-frequency components: $d^n U(\omega,t)/dt^n \propto \omega^n U(\omega,t)$, where $U(\omega,t)$, is the Fourier series of $u(t)$, and ω is the radial frequency.

Algorithms

There follows an algorithmic description of each of the methods considered. Before applying any of these methods, we remove the mean and if the signal contains long-period fluctuations, we remove these by high-pass filtering. The mean and the long-period fluctuations are added back in after despiking.

RC Filters Method. The name of this method comes from an equivalent electrical circuit that can produce the same result (Otnes and Enochson 1978). From the original time series u_i generate two time series $[\overline{u_i}]^2$ and $\overline{u_i^2}$, where $[\overline{u_i}]^2$ is the square of u_i after low pass filtering and $\overline{u_i^2}$ is the low pass filter of u_i^2. The sample variance σ_i^2 is

$$\sigma_i^2 = \overline{u_i^2} - [\overline{u_i}]^2 \tag{3}$$

The point $i+1$ is accepted as good if

$$\overline{u_i} - k\sigma_i < u_{i+1} < \overline{u_i} + k\sigma_i \tag{4}$$

where k=parameter, usually set between 3 and 9.

Tukey 53H Method. The Tukey 53H method uses the principle that the median is a robust estimator of the mean to generate a smooth time sequence that can be subtracted from the original signal:

1. Construct a sequence $u_i^{(1)}$ from the median of the five data points from u_{i-2} to u_{i+2};
2. Construct a sequence $u_i^{(2)}$ from the median of the three data points from $u_{i-1}^{(1)}$ to $u_{i+1}^{(1)}$;
3. Construct the Hanning smoothing filter $u_i^{(3)} = \frac{1}{4}(u_{i-1}^{(2)} + 2u_i^{(2)} + u_{i+1}^{(2)})$;
4. Construct the sequence $\Delta_i = |u_i - u_i^{(3)}|$ and reject the point if $\Delta_i > k\sigma$, where k is a predetermined threshold and σ is the standard deviation of u_i; and
5. Replace the spike.

Acceleration Thresholding Method. This method is a detection and replacement method with two phases: one for negative accelerations and the second for positive accelerations. In each phase, numerous passes through the data are made until all data points conform to the acceleration criterion $\lambda_a g$ and the magnitude threshold $k\sigma$. The steps in each phase are:

1. Calculate the acceleration from $a_i = (u_i - u_{i-1})/\Delta t$, where Δt is the sampling interval; and
2. Identify those points where $a_i < -\lambda_a g$ and $u_i < -k\sigma$ and replace them.

Step 2 is repeated until no more spikes are detected, then the second phase is begun:

1. Calculate the acceleration as above; and
2. Identify those points where $a_i > \lambda_a g$ and $u_i > k\sigma$ and replace them.

Step 2 is repeated until no more spikes are detected.

Experience shows that good choices for the parameters are: $\lambda_a = 1-1.5$ and $k = 1.5$.

Wavelet Thresholding Method. This method is a detection and replacement method that is similar to the previous method, except that all the calculations are done in wavelet space. The wavelet transform is analogous to the Fourier transform, except that the basis function (mother wavelet), instead of being a continuous cosine function, is a function that has compact support. In wavelet space, the signal becomes a series of coefficients, or details, relating to the position in time and degree of dilation of the wavelet. The wavelet thresholding method uses the lowest scale wavelet coefficient, often called the first detail, $d_{1,i}$ (e.g., Ogden 1997), calculated from the convolution of the signal with the mother wavelet with unit dilation and at various times i. There is a wide range of mother wavelets available, from the simple box shape (Haar) to the more complicated symmetric and asymmetric wavelets described in Daubechies (1992). If the Haar mother wavelet is chosen, this and the previous method reduce to almost the same algorithm, except that for the wavelet method the index i increments by two rather than one. Similar to the acceleration thresholding method, the data are passed through several times until no more spikes are detected. Before starting, the mean must be removed and to avoid end effects the beginning of the record needs to be padded out with a number of zeros equal to at least the number of iterations. The first data point is discarded after each iteration. In each iteration the steps are:

1. Calculate the first wavelet coefficient, $d_{1,i} = \int_{-\infty}^{\infty} u(t)\psi_{1,i}(t)dt$, where $\psi_{1,i}(t)$ is the mother wavelet centered at i, with unit dilation;
2. Identify points where $|d_{1,i}| > \lambda_U \hat{\sigma}$ (this threshold is discussed below);
3. Establish a sequence $\hat{d}_{1,i}$ of zeros except for the locations of spikes where $\hat{d}_{1,i} = 1$;
4. Calculate the inverse wavelet transform of $\hat{d}_{1,i}$ to yield a time series of zeros except for the points that are spikes (this identifies the locations in the time series where spikes exist); and
5. Replace the spikes.

There are a number of options for the threshold in Step 2. The one that is shown uses the Universal threshold, Eq. (2), from Donoho and Johnstone (1994). Katul and Vodakovic (1998) give two possibilities for the estimator $\hat{\sigma}$:

$$\hat{\sigma} = \sqrt{\frac{1}{n/2-1}\sum_{i=1}^{n/2}(d_{1,i} - \overline{d})^2} \tag{5}$$

where $\overline{d}$=mean of $\hat{d}_{1,i}$, or, as they state, a more robust value:

$$\hat{\sigma} = \frac{1}{0.6745}\langle |d_{1,i} - \overline{d}| \rangle_i \tag{6}$$

where $\langle \cdots \rangle_i$ denotes the mean over i.

Another alternative is to use the acceleration criterion from the previous method. An analysis of these options is included in the Results.

Phase-Space Thresholding Method. Here we introduce a new method that uses the concept of a three-dimensional Poincaré map or phase-space plot in which the variable and its derivatives are plotted against each other. The points are enclosed by an ellipsoid defined by the Universal criterion and the points outside the ellipsoid are designated as spikes. The method iterates until the number of good data becomes constant (or, equivalently, the number of new points identified as spikes falls to zero). Each iteration has the following steps:

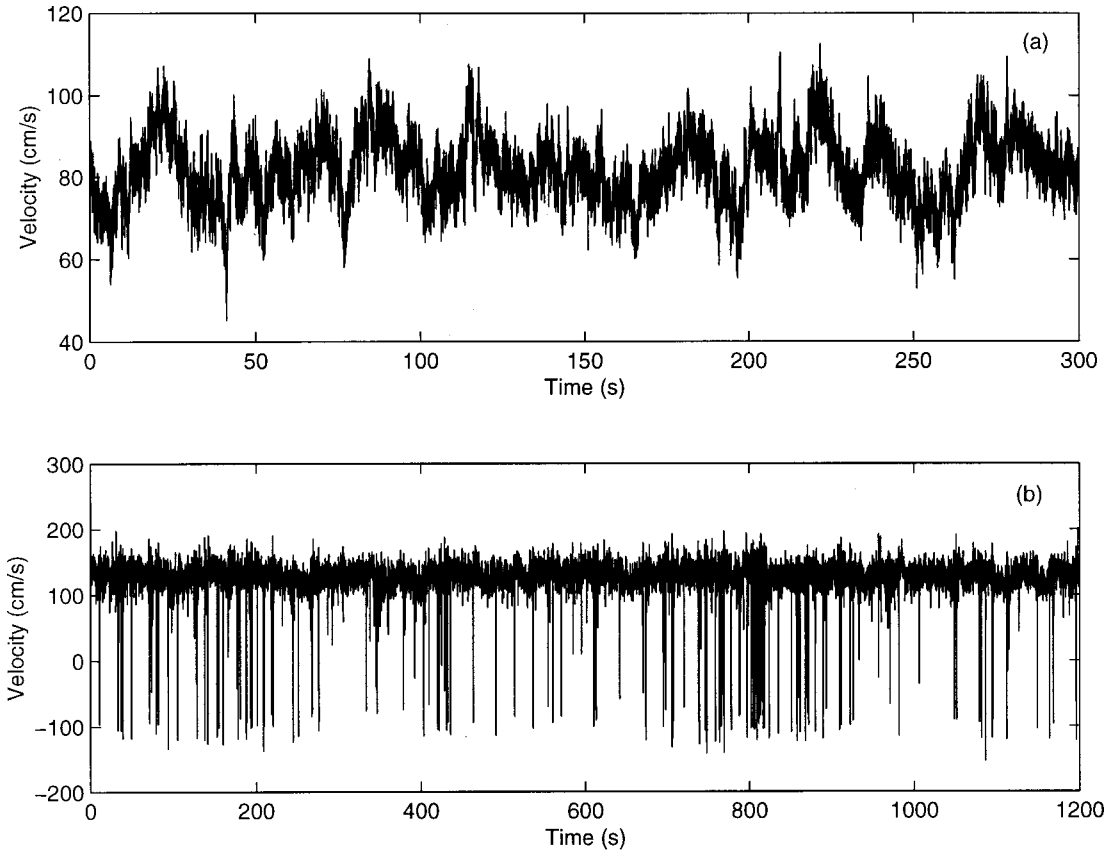

Fig. 3. Sample datasets used to test despiking algorithms: (a) clean record and (b) contaminated record

1. Calculate surrogates for the first and second derivatives from:

$$\Delta u_i = (u_{i+1} - u_{i-1})/2 \tag{7}$$

$$\Delta^2 u_i = (\Delta u_{i+1} - \Delta u_{i-1})/2 \tag{8}$$

(Note: we *do not* divide by the time step Δt—the reasons are given below.)

2. Calculate the standard deviations of all three variables, σ_u, $\sigma_{\Delta u}$, and $\sigma_{\Delta^2 u}$, and thence the expected maxima using the Universal criterion, Eq. (2).
3. Calculate the rotation angle of the principal axis of $\Delta^2 u_i$ versus u_i using the cross correlation:

$$\theta = \tan^{-1}\left(\sum u_i \Delta^2 u_i \Big/ \sum u_i^2\right) \tag{9}$$

(Note: for Δu_i versus u_i and for $\Delta^2 u_i$ versus Δu_i $\theta \equiv 0$ because of symmetry.)

4. For each pair of variables, calculate the ellipse that has maxima and minima from 3 above. Thus, for Δu_i versus u_i the major axis is $\lambda_U \sigma_u$ and the minor axis is $\lambda_U \sigma_{\Delta u}$; for $\Delta^2 u_i$ versus Δu_i the major axis is $\lambda_U \sigma_{\Delta u}$ and the minor axis is $\lambda_U \sigma_{\Delta^2 u}$; and for $\Delta^2 u_i$ versus u_i the major and minor axes, a and b, respectively, can be shown by elementary geometry to be the solution of

$$(\lambda_U \sigma_u)^2 = a^2 \cos^2\theta + b^2 \sin^2\theta \tag{10}$$

$$(\lambda_U \sigma_{\Delta^2 u})^2 = a^2 \sin^2\theta + b^2 \cos^2\theta \tag{11}$$

5. For each projection in phase space, identify the points that lie outside of the ellipse and replace them.

At each iteration, replacement of the spikes reduces the standard deviations calculated in 2 and thus the size of the ellipsoid reduces until further spike replacement has no effect.

Care must be taken in the calculation of θ to ensure that Eqs. (10) and (11) do not become ill conditioned. This can occur if σ_u and $\sigma_{\Delta^2 u}$ are orders of magnitude different. For example, if $\sigma_u \gg \sigma_{\Delta^2 u}$, then θ will be small, but $a^2 \sin^2\theta$ may be significant in comparison to $b^2 \cos^2\theta$ and the solution of Eqs. (10) and (11) may be complex. A solution to this problem is to refrain from dividing the numerical derivatives by the time step in Step 1. This yields velocity differences that are of the same order as the velocity. It also means that all the axes have the same units ($\mathrm{cm\,s^{-1}}$ for ADV data).

Spike Replacement

Once a spike has been detected, the problem becomes one of deciding what to replace it with. There are numerous alternatives:

1. extrapolation from the preceding data point: $u_i = u_{i-1}$;
2. extrapolation from the two preceding points: $u_i = 2u_{i-1} - u_{i-2}$;
3. the overall mean of the signal;
4. a smoothed estimate; or
5. interpolation between the ends of the spike.

Of these, the smoothed estimate is the most aesthetically pleasing, but it has no more validity than the others. Extrapolation from the preceding one or two points is especially efficacious for the acceleration thresholding method where we are passing through the data detecting spikes ahead of us and replacing them. However, for turbulence data the use of the two preceding points can produce wild extrapolations. Alternatively, using just one preceding point can produce deep, wide troughs if the spike has multiple points. Using the overall mean of the signal solves this problem, but has the disadvantage that the replacement can introduce an-

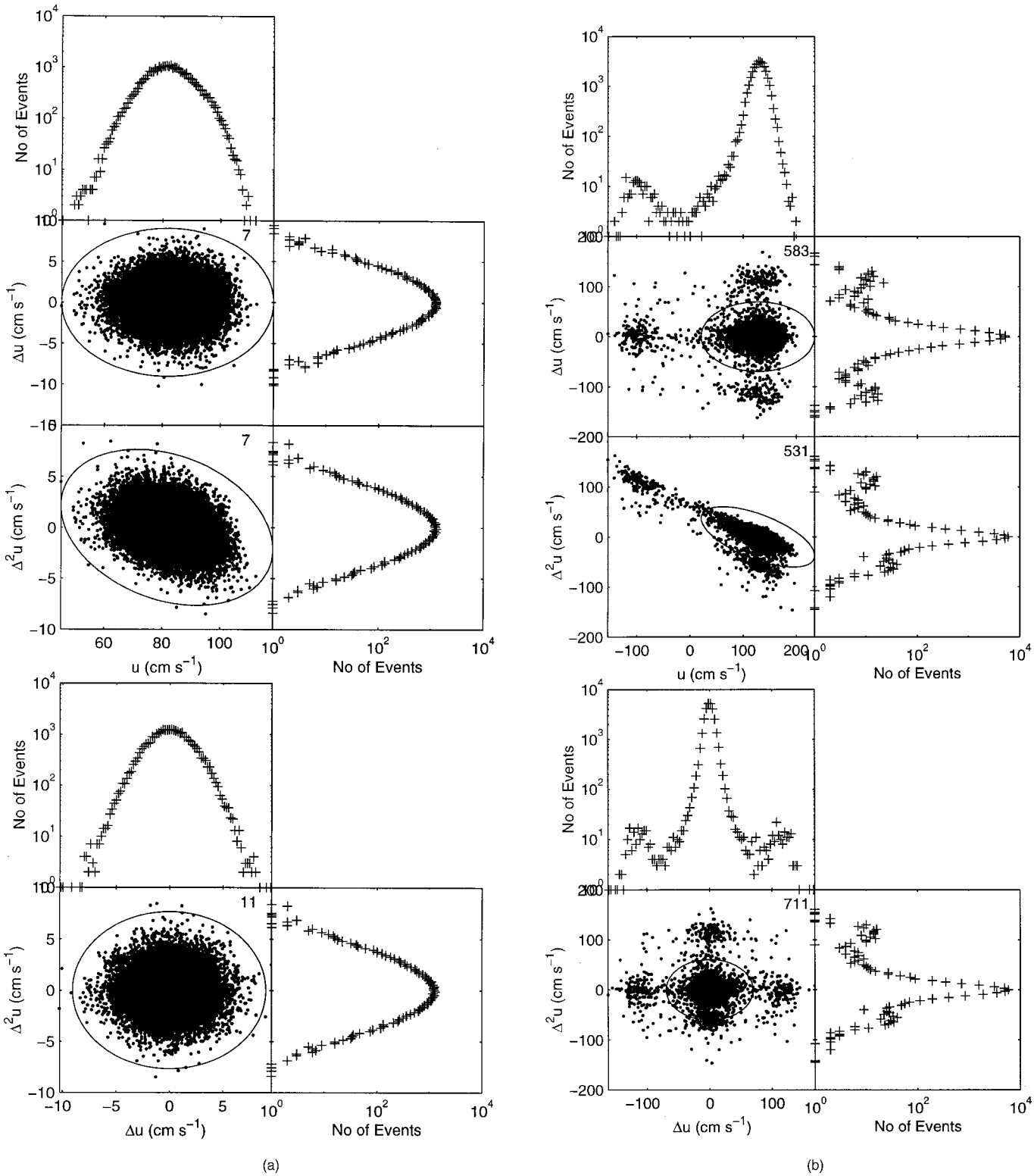

Fig. 4. Phase-space plots for (a) clean dataset and (b) contaminated dataset (number of spikes detected are listed in top right corner of each phase-space plot)

other spike if there is a local departure from the mean. Smoothed estimates arise naturally from the *RC* filters and Tukey 53H methods and are a byproduct of the wavelet method (from the remainder obtained by subtracting the inverse transform of the first wavelet coefficient from the signal). However, the smoothed estimates from these methods contain spurious data from the spike itself. Interpolation across the spikes works well in most cases, providing the spike starts and finishes at about the same level; but if the level at the start is substantially different from the level at the finish, a straight-line interpolation may generate an additional spike that can be detected, but not replaced. After much searching, the most satisfactory method we have found is to use a poly-

Table 1. Results of Applying Five Despiking Methods to Clean Data set for Various Parameters

Method	Parameters	No. of spike events
RC filters[a]	$k=6$	474
	$k=9$	187
Tukey 53H	$k=1$	174
	$k=1.5$	13
Acceleration	$\lambda_a=1$, $k=1.5$	91
	$\lambda_a=1.5$, $k=1.5$	5
	$\lambda_a=1$, $k=2$	46
Wavelet	$\lambda_a=1.5$	10
	Universal	0
Three-dimensional phase space	Universal	9

[a]Butterworth low-pass filter of fifth order with a frequency cutoff at 10% of the Nyquist frequency was used (Anonymous 1996).

nomial of best fit through the data on either side of the spike, then interpolate across the spike with this polynomial. Trials have shown that for ADV data the best options are to use a third-order polynomial (i.e., a cubic) through 12 points on either side of the spike. A cubic allows enough curvature without introducing extra spikes. The number of points for fitting needs to be large for ADV data to ensure that local, representative levels are found on either side of the spike. The corollary is that there must not be any spikes within the 12 data points on either side of the spike under consideration. This means that some spike events have to be amalgamated, reducing the number of events, but making some of them wider. Of course, the number of points used in fitting the polynomial may depend upon the sampling frequency, but we have found that 12 is suitable for sampling rates in the range from 25 to 100 Hz that we routinely use.

Results

Tests were carried out on all five methods to determine the method that best passes a ''clean'' sequence unchanged, but will clean up a contaminated data set. We chose, as an example, the clean and contaminated data sets shown in Fig. 3 from our vast library of ADV measurements. In the normal course of events, the contaminated data set would have been rejected after visual inspection as having too many spikes, so it is a pathological case. Nevertheless, it provides ample opportunities to test the algorithms. In particular, we need to assess the capability of each method to detect double-point spikes such as that shown in Fig. 1 at about 39 s. We expect that all the algorithms will be able to handle single-point spikes such as that shown in Fig. 1 at about 37 s.

The distributions of the test data are presented in Figs. 4(a and b) [corresponding to Figs. 3(a and b), respectively]. Each figure contains the three projections of the three-dimensional phase space (Fig. 2) with the individual data plotted as points and the ellipse defined by the Universal thresholds plotted as a continuous curve. The number of points lying outside each ellipse is printed on the upper right of each phase-space plot. Above and alongside each phase-space projection is the corresponding histogram showing the distribution of points.

We would expect that very few spikes would be found in the clean record and Table 1 shows to what extent this is true for the despiking methods under consideration. In fact, all of them perform adequately, except for the *RC* filters method that identifies far too many spikes, no matter how large we make parameter *k*. An additional difficulty with this method is that its performance depends to a large extent on the choice of low-pass filter. The Tukey 53H method is quite sensitive to the choice of the parameter *k*, and the acceleration method is more sensitive to the acceleration threshold than to the velocity threshold. This sensitivity to parameters is a disadvantage of these methods and highlights the advantage of methods that use the Universal threshold which requires no parameters.

On the other hand, the wavelet method appears to be less sensitive to the choice of parameters, with the acceleration criterion using $\lambda_a=1.5$ performing almost as well as the Universal threshold (Table 1). The reason for this is that the localized nature of the mother wavelet exaggerates the spike when it is transformed to wavelet space, so that it stands out above the surrounding data.

The real test of the methods is application to a contaminated record like that in Fig. 3b. Table 2 lists the number of spikes detected by each method, using the optimum parameters from Table 1. The first point to note is that in spite of the spiky appearance of Fig. 3b, the number of spikes is small compared to the total number of data ($<3\%$). We also need to point out that while most of the methods identify individual spikes, the wavelet and phase-space methods identify spike events. In the case of the wavelet method, these are all two-point events, and for the phase-space methods the events vary in length from 1 point up to 15 points. Fig. 5 shows the way each method handles a typical, complex, multipoint spike event. All of the methods detect the two deep spikes, but only the acceleration and the phase-space methods eliminate the two points between the spikes and the shallow spike that follows the two deep spikes. In fact, for the phase-space

Table 2. Results of Applying Five Despiking Methods to Contaminated Data set with Optimum Parameters from Table 1, Where N_{it} Is Number of Iterations

Despiking method	Parameters	No. of spike events	N^{it}	Final standard deviation ($cm\,s^{-1}$)	Replacement strategy
RC filters[a]	$k=9$	253	2	18.33	1 preceding
Tukey 53H	$k=1.5$	689	3	14.04	1 preceding
Acceleration	$\lambda_a=1.5$, $k=1.5$	834	10[b]	13.17	1 preceding
Wavelet	Universal	213	2	14.33	Mean
Three-dimensional phase space	Universal	194	4	13.78	Cubic fit
Original signal	...	...	...	23.52	...

[a]Butterworth low-pass filter of fifth order with a frequency cutoff at 10% of the Nyquist frequency was used (Anonymous 1996).

[b]Five iterations for deceleration and five for acceleration.

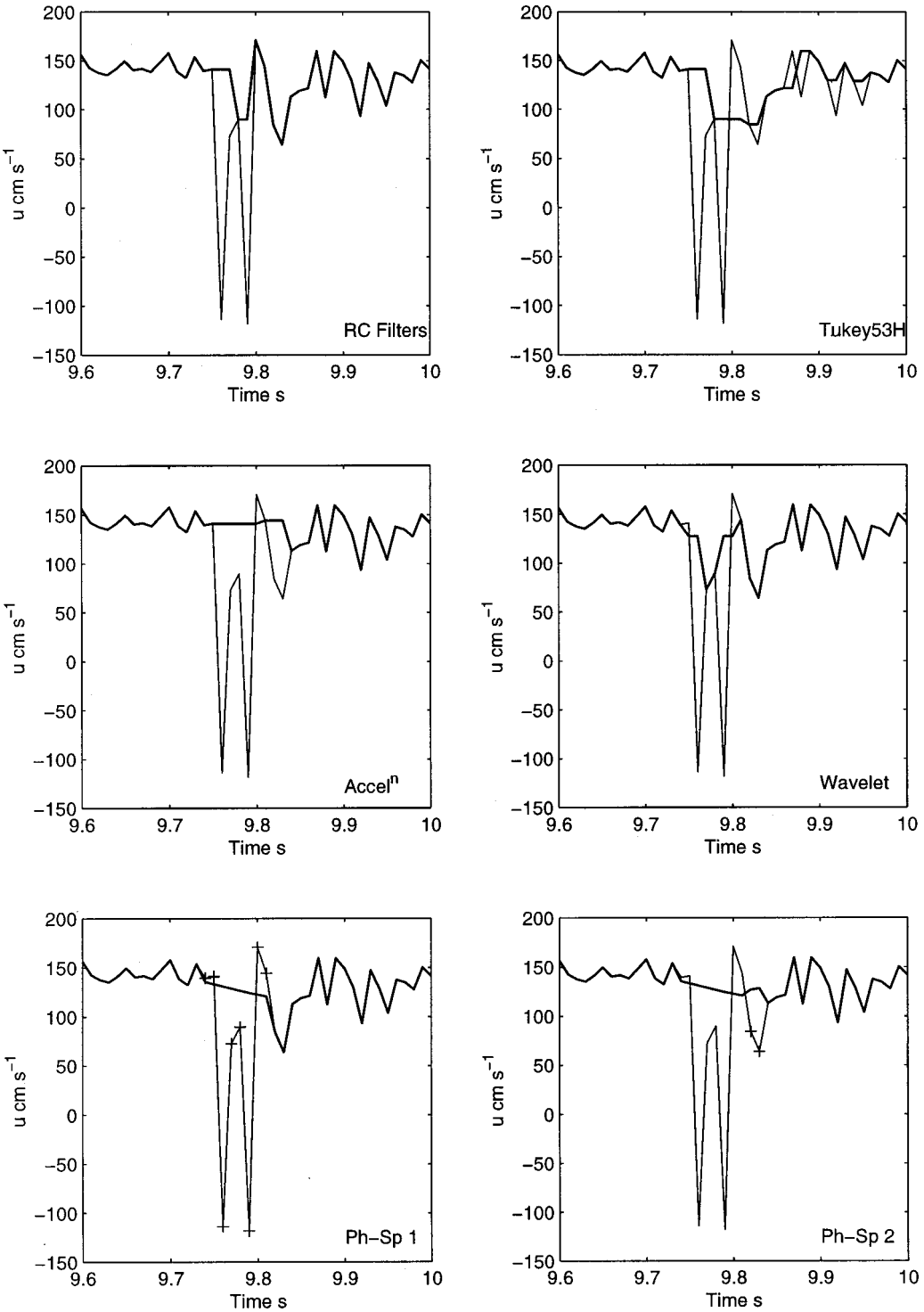

Fig. 5. Detection of typical multipoint spike event (light line) and its replacement (heavy line) using various methods and parameters in Table 2

method, the shallow spike is only detected on the second iteration (Ph-Sp 2 in Fig. 5) and this causes a problem because the data from this spike are used in the first iteration to fit the cubic that is used for interpolating over the spike event. Thus, when the shallow spike is detected in the second iteration, the data used for fitting the cubic for interpolation over this event are in fact the data that were fitted in the previous iteration. This causes the small rise in velocity in the replaced points. A similar, but in this case less severe, problem occurs in the acceleration method when the replacement strategy of using the preceding point is used. These problems arise because of the length of the spike event (8 points) and the proximity of the next event (the next 2 points). As mentioned earlier, this is a pathological case.

Table 2 also lists the standard deviation of the final despiked record in each case. These should be compared with the standard deviation of the original signal of 23.52 cm s^{-1}. Notice that the standard deviation from the acceleration method is less than that from the phase-space method. The reason is that the acceleration

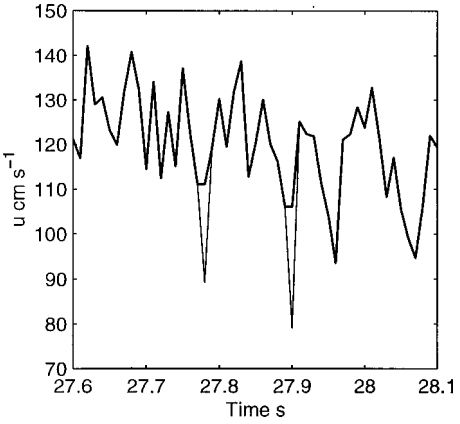

Fig. 6. Two spikes that are detected and replaced by acceleration method, but not by phase-space method

method has identified some points as spikes that may not be spikes at all, such as are shown in Fig. 6. This illustrates a problem with the acceleration method that we have had difficulty resolving, namely: how do we set the two parameters so as to eliminate spikes, but not damage the good data?

Finally, in Fig. 7 we present the phase-space projections of the despiked, contaminated record and in Fig. 8 we show the despiked record itself. The difference between Fig. 7 and Fig. 4(b) for the data before despiking is remarkable. Whereas in Fig. 4(b) there are a large number of points outside the ellipsoid and the histograms have secondary peaks, in Fig. 7 almost all the data lie within the ellipsoid. In the phase-space plot of $\Delta^2 u_i$ versus Δu_i, the string of velocities well below the mean value and having zero second derivatives corresponds to a single spike event similar in appearance to that in Fig. 5, but comprising 26 points resulting from the amalgamation of adjacent spike events. So, despiking identifies these as spikes, but they cannot be improved because their replacements would continue to be ''spikes'' in the next iteration. Thus, while the despiking has cleaned the record significantly, and the resulting time series looks perfectly satisfactory (Fig. 8), there are remaining features that cannot be accommodated. These are unlikely to cause problems in the calculation of low-order statistical moments, but care must be exercised when calculating high-order moments or mulitfractals from such data, and as a matter of routine we would reject this record for that reason.

Discussion

The results show that the phase-space thresholding method works extremely well and this has been confirmed by successful application of the method to numerous ADV data sequences from our data archive. Given that such data are neither independent (they have statistically significant autocorrelations at lag$=1$) nor normally distributed (the tails of the frequency distribution decay slower than normal) and therefore violate the basic assumption in the derivation of the Universal threshold, the question is: why does it work so well? The answer is twofold. First, the velocity spectrum at low frequencies approximates the white-noise spectrum (Nikora and Goring 2000) and therefore providing the data sequence is long enough to encompass the maximum external time scale of the application, the Universal threshold applies. If the data sequence were undergoing a sustained change in velocity

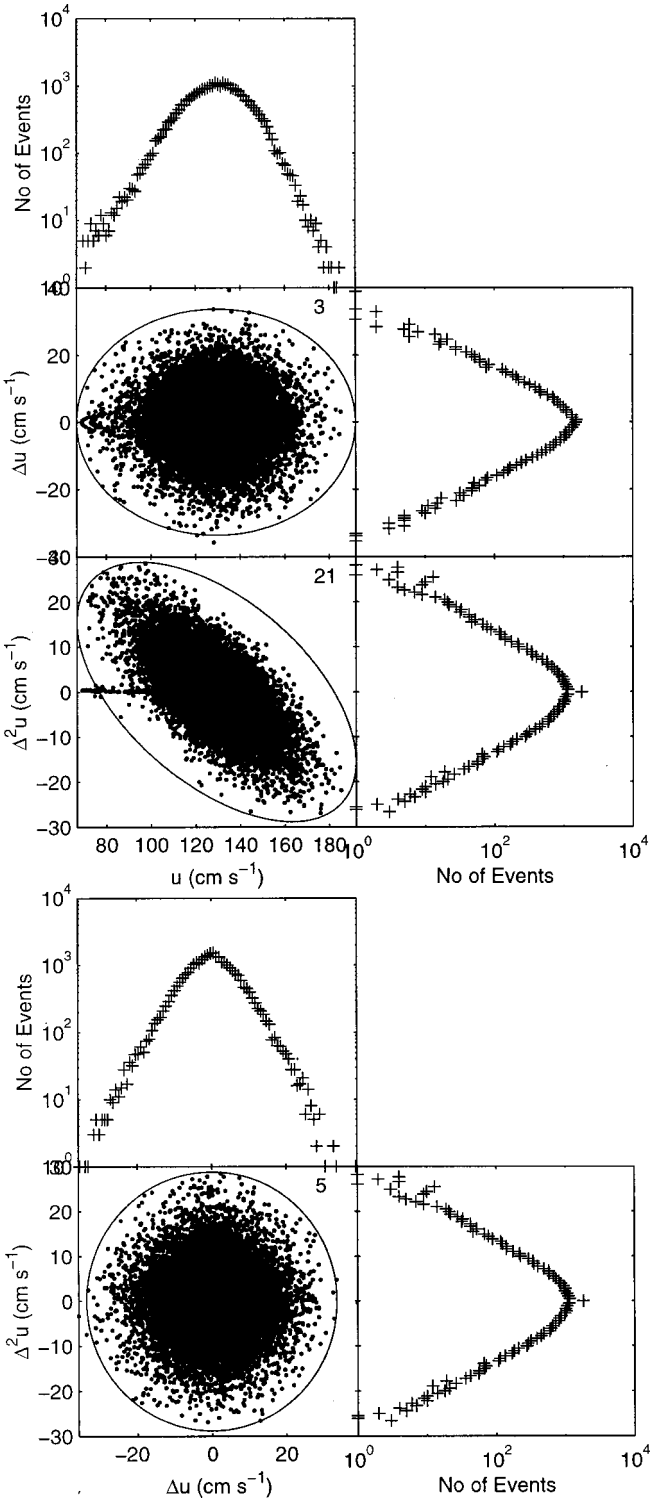

Fig. 7. Phase-space plots of despiked contaminated data set (number of spikes detected in last iteration are listed in top right corner of each phase-space plot)

(e.g., over a tidal cycle or during a change in river flow), then the phase-space method would not work unless these long-scale fluctuations were removed by high-pass filtering. Second, providing the data are approximately normal, the low-order moments, specifically the standard deviation, are adequate to describe the distribution.

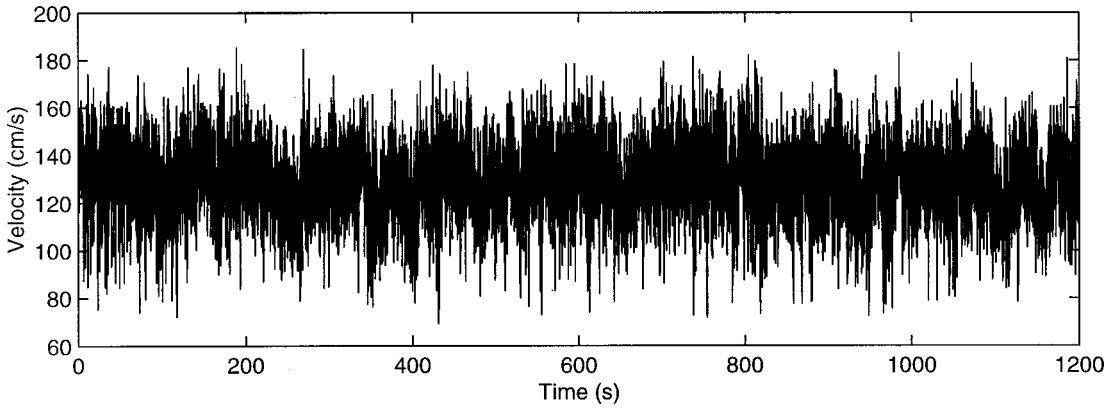

Fig. 8. Contaminated record of Fig. 3(b) after cleaning using phase-space method and replacing each spike with cubic polynomial fitted to 12 points on either side of spike

Numerical experiments undertaken to test the Universal threshold revealed that the threshold is about 10% high for $n > 1{,}000$. Initially, we thought this was either a problem with our random number generator or that the number of samples was insufficient to arrive at a valid result. Therefore, we undertook some tests on a massively parallel computer using the scalable, parallel random-number generator (SPRNG) of Mascagni et al. (1999) to generate uniformly distributed pseudorandom numbers and the well-known Box–Muller algorithm (Box and Muller 1958) to convert those into a normal distribution. For each number of data n we ran 1028 Monte Carlo simulations using 128 processors, thus generating $2^{17} = 131{,}072$ simulations for each n. The tests confirmed that the Universal threshold is 10% high. For despiking, we believe this bias in the threshold is not really a problem. After all, our data are only approximately normal in their distribution. Nevertheless, it does mean that fewer spikes are detected than if the Universal threshold were exact.

ADVs measure velocity along three beams, then convert these to Cartesian coordinates using a matrix transformation. Therefore, we would expect that a spike detected in one of the transformed velocities would also occur in the others. However, our experience is that for a down-looking ADV, the vertical velocity has many fewer spikes than the horizontal velocities. Therefore, our procedure is to despike each velocity component separately, and to record the number of spikes detected.

Conclusions

In this note we have presented several different techniques for detecting and replacing spikes in ADV data sequences. Single-point spikes that protrude above or below the surrounding data are easily detected and replaced. However, multipoint spikes and spikes that blend with the background are much more difficult to detect. Of the methods considered, the phase-space thresholding method is the most suitable for detecting spikes in these data. The method is new, but comprises a combination of three concepts that are not new: (1) that differentiation enhances the high-frequency components of a signal; (2) that the expected maximum of a sequence of random numbers is given by the Universal threshold $\sqrt{2 \ln n}\sigma$, and (3) that good data cluster in a dense cloud in three-dimensional phase space. It is also the most robust because it requires no external parameters. Other methods such as the one based on an acceleration criterion suffer from ambiguity in what thresholds to choose and the fact that different records may need different thresholds.

Spike replacement is an arbitrary procedure. There are many different strategies available and none of them has more validity than any other. We have found that for ADV data with sampling frequencies from 25 to 100 Hz, the best solution is to use 12 points on either side of the spike to fit a third-order polynomial that is interpolated across the spike.

Acknowledgments

The research was conducted under Contract Nos. CO1X0023 and CO1X0024 from the Foundation of Research, Science and Technology (New Zealand) and Contract No. NIW701 (Intermittency and Bursting Phenomena) from the Marsden Fund administered by the Royal Society of New Zealand. The writers gratefully acknowledge the assistance of our technician Frazer Munro for data preparation and algorithm testing. Three anonymous reviewers provided constructive comments that we have gratefully incorporated into the manuscript.

Notation

The following symbols are used in this paper:

a,b = major and minor axes, respectively, of projection of ellipsoid on plane of $\Delta^2 u_i$ versus u_i;
a_i = discrete acceleration using backward differences;
$d_{1,i}$ = first wavelet coefficient of u_i;
$\hat{d}_{1,i}$ = first wavelet coefficient of spikes (1 = a spike, 0 otherwise);
$E(..)$ = expected value of;
g = acceleration of gravity;
k = standard deviation threshold;
n = number of data;
t = time;
$U(\omega,t)$ = Fourier series of $u(t)$;
$u(t)$ = velocity time series;
u_i = discrete velocity time series;

Δ_i = difference between rough and smooth datasets;
Δt = time interval between data points;
Δu_i = surrogate for first derivative of u_i using central differences;
$\Delta^2 u_i$ = surrogate for second derivative of u_i using central differences of Δu_i;
θ = angle of rotation of principal axis of $\Delta^2 u_i$ versus u_i;
λ_a = acceleration threshold;
λ_U = Universal threshold;
ξ_i = independent, identically distributed, standard, normal, random variable;
σ = standard deviation;
$\hat{\sigma}$ = estimate of standard deviation;
$\sigma_u, \sigma_{\Delta u}, \sigma_{\Delta^2 u}$ = standard deviations of $u_i, \Delta u_i, \Delta^2 u_i$, respectively; and
ω = radial frequency.

References

Abarbanel, H. D. I. (1995). *Analysis of observed chaotic data*, Springer, New York.

Addison, P. S. (1997). *Fractals and chaos: an illustrated course*, IOP, London.

Anonymous. *Matlab signal processing toolbox user's guide*. (1996). The MathWorks Inc.

Box, G. E. P., and Muller, M. E. (1958). ''A note on the generation of random normal deviates.'' *Ann. Math. Stat.*, 29, 610–611.

Daubechies, I. (1992). *Ten lectures on wavelets*, SIAM, Philadelphia.

Donoho, D. L., and Johnstone, I. M. (1994). ''Ideal spatial adaptation by wavelet shrinkage.'' *Biometrika*, 81(3), 425–455.

Katul, G. G., and Vodakovic, B. (1998). ''Identification of low-dimensional energy containing flux transporting eddy motion in the atmospheric surface layer using wavelet thresholding methods.'' *J. Atmos. Sci.*, 55, 377–389.

Mascagni, M. Ceperley, D., and Srinivasan, A. (1999). ''SPRNG: a scalable library for pseudo-random number generation.'' *Proc., 3rd Int. Conf. on Monte Carlo and Quasi Monte Carlo Methods in Scientific Computing*, Springer, Berlin.

McKinney, J. P. (1993). ''A method for locating spikes in a measured time series.'' *Proc., 2nd Int. Symposium on Ocean Wave Measurement and Analysis*, 338–393.

Nikora, V. I., and Goring, D. G. (1998). ''ADV measurements of turbulence: can we improve their interpretation?'' *J. Hydraul. Eng.*, 124(6), 630–634.

Nikora, V. I., and Goring, D. G. (2000). ''Flow turbulence over fixed and weakly mobile gravel beds.'' *J. Hydraul. Eng.*, 126(9), 679–690.

Nikora, V. I., Goring, D. G., and Biggs, B. J. F. (1998). ''Silverstream eco-hydraulics flume: hydraulic design and tests.'' *NZ J. Marine Freshwater Res.*, 32(4), 607–620.

Ogden, R. T. (1997). *Essential wavelets for statistical applications and data analysis*, Birkhauser, Boston.

Otnes, R. K., and Enochson, L. (1978). *Applied time series analysis*, Wiley, New York, Vol. 1.

Roy, M., Kumar, V. R., Kulkarni, B. D., Sanderson, J., Rhodes, M., and van der Stappen, M. (1999). ''Simple denoising algorithm using wavelet transform.'' *J. Am. Ins. Chem. Eng.*, 45(11), 2461–2466.

Voulgaris, G., and Trowbridge, J. H. (1998). ''Evaluation of the acoustic Doppler velocimeter (ADV) for turbulence measurements.'' *J. Atmos. Ocean. Technol.*, 15(1), 272–289.