Assessment of Despiking Methods for Turbulence Data in Micrometeorology

DEREK STARKENBURG

Geophysical Institute, and Department of Atmospheric Sciences, College of Natural Science and Mathematics, University of Alaska Fairbanks, Fairbanks, Alaska

STEFAN METZGER

Fundamental Instrument Unit, National Ecological Observatory Network, and Institute of Arctic and Alpine Research, University of Colorado Boulder, Boulder, Colorado

GILBERTO J. FOCHESATTO

Geophysical Institute, and Department of Atmospheric Sciences, College of Natural Science and Mathematics, University of Alaska Fairbanks, Fairbanks, Alaska

JOSEPH G. ALFIERI

Hydrology and Remote Sensing Laboratory, ARS, USDA, Beltsville, Maryland

RUDIGER GENS, ANUPMA PRAKASH, AND JORDI CRISTÓBAL

Geophysical Institute, University of Alaska Fairbanks, Fairbanks, Alaska

(Manuscript received 28 July 2015, in final form 14 January 2016)

ABSTRACT

The computation of turbulent fluxes of heat, momentum, and greenhouse gases requires measurements taken at high sampling frequencies. An important step in this process involves the detection and removal of sudden, short-lived variations that do not represent physical processes and that contaminate the data (i.e., spikes). The objective of this study is to assess the performance of several noteworthy despiking methodologies in order to provide a benchmark assessment and to provide a recommendation that is most applicable to high-frequency micrometeorological data in terms of efficiency and simplicity. The performance of a statistical time window–based algorithm widely used in micrometeorology is compared to three other methodologies (phase space, wavelet based, and median filter). These algorithms are first applied to a synthetic signal (a clean reference version and then one with spikes) in order to assess general performance. Afterward, testing is done on a time series of actual CO_2 concentrations that contains extreme systematic spikes every hour owing to instrument interference, as well as several smaller random spike points. The study finds that the median filter and wavelet threshold methods are most reliable, and that their performance by far exceeds statistical time window–based methodologies that use the median or arithmetic mean operator (−34% and −71% reduced root-mean-square deviation, respectively). Overall, the median filter is recommended, as it is most easily automatable for a variety of micrometeorological data types, including data with missing points and low-frequency coherent turbulence.

1. Introduction

It is well known that the evaluation of micrometeorological turbulent fluxes depends on data quality, especially if cross comparisons between different experimental sites are performed, or if a quantification of natural variability is desired (Papale et al. 2006; Mauder et al. 2013). Differences in instrumentation and data handling/processing methods can cause larger differences in derived meteorological values than the variability of the natural processes to be quantified (Foken and Wichura 1996; Mauder and Foken 2006). Mauder et al. (2013) provide a comprehensive assessment of the variety of problems that can affect scalar flux calculations, beginning with the quality of the raw high-frequency data,

Corresponding author address: Gilberto J. Fochesatto, Geophysical Institute, University of Alaska Fairbanks, 903 Koyukuk Drive, Fairbanks, AK 99775-7320.
E-mail: gjfochesatto@alaska.edu

DOI: 10.1175/JTECH-D-15-0154.1

followed by the processing and assessment of derived flux values based on flow conditions. One fundamental problem with raw data is the presence of spikes, which are unrealistically high local deviations of one or a few data points compared to neighboring values (Vickers and Mahrt 1997). Spikes may manifest as artifacts of the instrument (Brock 1986), such as from water on the sonic anemometer transducers (Vickers and Mahrt 1997; Papale et al. 2006), from animal disturbances (Mauder et al. 2013), or from an unreliable power supply (Lee et al. 2004; Mauder et al. 2013). Other types of signal error such as dropouts (periods of data abruptly offset from the local mean), or sensor errors like calibration drift and hysteresis (Brock 1986), are often more visually evident. Spikes, however, are ubiquitous, and therefore an automated method for their removal is ideal. Spikes are problematic because they occupy the higher frequencies of a time series, but their presence affects all spectral characteristics of the data (Brock 1986). Even if spikes do not always have a significant effect on derived fluxes (Mauder and Foken 2006), they substantially affect higher-order statistical moments such as variance (Lee et al. 2004), which are required, for example, in applications such as source attribution models. Specifically, Fratini and Mauder (2014) attribute order-of-magnitude differences in variance to differences in despiking methods.

Furthermore, in the analysis by Vickers and Mahrt (1997, p. 513), testing for spikes is "the first quality control test and the only test that modifies the data itself." Spike detection is also part of the first stage in the algorithm for assessing the uncertainty of calculated flux values described by Mauder et al. (2013), and the second step after verification of physically plausible values in the eddy covariance software package TK3 by Mauder and Foken (2011). As such, despiking is the forerunner of basic data quality issues, and its performance sets the stage for subsequent quality control procedures and/or data usage. A despiking algorithm will ultimately flag a certain number of spiked points that can then be interpolated (i.e., replaced via an estimation of some sort), or marked as missing values. The former will directly control the number of artificial data points entered into the analysis; the latter will indirectly affect the data analysis by establishing a number of missing points, an amount of which may result in the data being rejected (Vickers and Mahrt 1997; Mauder et al. 2013). Despiking is a powerful tool that requires cautious control in order to limit its aggressiveness on observed data that were acquired at great time and monetary expense. Therefore, we focus specifically on the performance of despiking algorithms in the present study.

Technological advancements in equipment and data quality correction methods have significantly improved to date (Foken et al. 2010, and citations therein). Many instruments now have internal capabilities to flag suspect data (Mauder et al. 2013), and quality control algorithms are becoming more standardized (Papale et al. 2006) and combined into software suites (Mauder and Foken 2011). As algorithms become readily available, they should not be treated as black boxes. Since instrumentation and site conditions vary, data cleaning algorithms should be tested before their universal application to diverse datasets. To this end, Papale et al. (2006) showed applicability of their proposed method of standardized net ecosystem exchange flux corrections to 12 annual datasets taken from eight European sites. They showed that the percentages of half-hourly data removed due to the despiking component of their analysis is quite similar across sites, which indicates that the algorithm is not likely biased by regionally specific characteristics of the data (see Table 2 in Papale et al. 2006).

One major problem with evaluating the performance of a data cleaning algorithm is that outside of obviously unphysical values, one cannot know a priori how incorrect a given set of data is. To contend with this, it is proposed to initially test several despiking methods by applying them to both clean and spiked versions of a synthetic signal, for which one has control over the exact number and location of spike points. Afterward, testing the algorithms on real data embedded with obvious spiked points will pave the way for parameterizing their general performances on turbulent micrometeorological data. In practice, a myriad of signals under a variety of conditions should be evaluated before long-term performance of these methods can be stated with confidence.

2. Despiking methodologies

Table 1 shows a classification of different despiking algorithms, along with some of their pros and cons, as discussed in the literature. These classifications are 1) a statistical approach based on the mean or median, 2) a filter approach using a least squares polynomial filter or a nonlinear median filter, 3) a phase-space approach in which a signal is compared with its first and second derivatives, and 4) a spectral approach using wavelet transformation. The subsections below provide detailed descriptions of these methodologies with a focus on the methods implemented in this paper.

a. Statistical-based methods

A popular data cleaning method based on Højstrup (1993), and further developed by Vickers and Mahrt (1997), makes use of the mean (μ) and standard deviation (σ) for the data within a window whose width in time (L) is 5 min, or 3000 measurements at 10 Hz. The window slides point by point, searches for values that

TABLE 1. Classification of despiking algorithms found in the literature. References in bold and italic fonts represent methods (with abbreviation) that are the subject of this study, or are used for intercomparison, respectively. The asterisk (*) denotes the most common method used to despike eddy covariance data.

Method	Reference	Procedure	Advantages	Disadvantages
Statistical				
Mean	Højstrup (1993)	Uses mean and standard deviation	Conceptually simple	
	***VM**			Requires parameterization and calibration
	Mauder and Foken (2011)			Requires iteration, thus computationally expensive
	Fratini and Mauder (2014)			
Median	*Metzger et al. (2012)*	Uses MAD	Median and MAD less affected by outliers than mean and standard deviation	Requires parameterization and calibration
				Requires iteration, thus computationally expensive
	Mauder et al. (2013)	Uses MAD	Median and MAD less affected by outliers than mean and standard deviation	Global threshold can be insensitive
			No iterations required	
			Not dependent on window length (L)	
Filter				
Polynomial	Phillips and Harris (1990)	Uses Savitsky–Golay filter over raw data	Filter preserves second-order statistical moments	
	Hill and Rogalia (1992)			Requires second evaluation after initial spikes are detected (i.e., first pass biased by initial spikes)
	Fochesatto and Sloan (2008)			
Median	**BR**	Uses median filter over raw data	Not affected by spike frequency or amplitude	Cannot detect more than N consecutive spikes (where $2N + 1$ = window length)
			Median less affected by outliers than mean	Window length must be reasonably small to best evaluate local data
			Conceptually simple	
Phase space				
	GN	Compares signal to its first two derivatives	Can detect multiple spike events	Requires iterations
			Does not require parameters for streamflow data (i.e., uses universal threshold)	May be overly sensitive to second derivatives for micrometeorological data
				Hard to gap fill if multiple neighboring spikes exist
				Requires removing mean and detrending
Spectral				
Donoho and Johnstone (1994)		Wavelet shrinkage (denoising)	Wavelet kernel can evaluate signal changes (frequency) while preserving location (time)	Is biased by missing data points
GN (WT)		WT (despiking)	Wavelet kernel can evaluate signal changes (frequency) while preserving location (time)	Is biased by missing data points

exceed $\pm 3.5\sigma$, and replaces them by interpolation. The scheme is repeated until no more spikes are detected, with the threshold of spike detection increasing by $\pm 0.1\sigma$ each time. Cases where four or more consecutive points exceed the threshold are not replaced. More recently, the algorithm by Vickers and Mahrt (1997) has been modified by Mauder and Foken (2011) for use in their eddy covariance software package, TK3. The algorithm by Vickers and Mahrt (1997) is also the basis for the despiking and statistical screening procedures in the

widely used software package EddyPro 3.0 (LI-COR 2012).

Other similar statistical approaches use the deviation of points from the median rather than from the mean, since the median is unbiased by outliers. Specifically, Metzger et al. (2012) apply the median absolute deviation (MAD), where the differences between individual points and a windowed (local) median ±MAD are used to locate spikes. They found that this method reliably removed spikes that were not detected by the Vickers and Mahrt (1997) and Mauder and Foken (2011) schemes. The median- and MAD-based method has subsequently been adopted by Mauder et al. (2013) in an update to their TK3 eddy covariance software package, which uses a nonwindowed (global) median ± MAD threshold. Because of superior performance, the median- and MAD-based method has also been proposed for EddyPro (Fratini and Mauder 2014).

This study will assess the performance of a version of the original scheme by Vickers and Mahrt (1997, hereafter VM), in order to provide a base-level comparison from which to evaluate other lesser-known methodologies. In this paper, VM will follow the same concept as what is described above from the original paper by VM, except that the window size is 300 s, or 6000 points for our 20-Hz data. VM will be allowed to iterate until no more spikes are detected, with spikes replaced by linear interpolation between successive iterations. Other statistical methods from Table 1 are not investigated in similar detail, but their overall performance is nonetheless quantified (section 2e) and a comparative summary of results for those other methods is provided for context (section 4b).

b. *Filter methods*

Taking advantage of the fact that the median is an unbiased statistical estimator, the paper of Brock (1986) carefully outlines a methodology whereby a nonlinear median filter is applied with threshold logic to remove spikes. This idea derives from Beaton and Tukey (1974), who were concerned with handling bad or missing lines in band-spectroscopic data; specifically, they noted that a smoothing filter that is nonlinear will not be biased by the spikes themselves. Nonlinear median filtering is a simple operation whose only parameter is the window size, and whose physical behavior is to maintain (''pass'') signal edges but remove (''stop'') impulses or outliers (Gallagher and Wise 1981). Based on the paper by Pasian and Crise (1984), Brock (1986) describes how a median filter can perform spike detection if a threshold logic is used to determine which data values are far enough from the filter to be considered outliers. Brock (1986) first produces a histogram of the differences between the raw signal and the median filtered signal (D_I). In such a distribution, valid data will condense into a central population, while spikes result in subpopulations. According to Brock (1986, p. 55), ''searching the histogram for the first minima from the center'' will define a threshold value (DT), and any points for which $|D_I| > \mathrm{DT}$ are considered spikes. The first minima in the distribution of D_I are ideally bins with zero counts (i.e., gaps); if no zero bin counts are found, then the bin size is doubled (e.g., from an initial size of 25–51, with one added to retain an odd number bin size) (Brock 1986). However, Brock (1986, p. 55) notes that, ''If the bin size is too small, artificial gaps can form because the data have finite resolution. . . .'' Thus, for histograms generated purely from acquired data, the maximum number of bins should be such that the bin width is not less than the acquiring instrument's resolution. For a histogram of differences (D_I) as in this case, the restriction is the maximum number of bins, which cannot exceed the digital resolution of the instrument used to calculate D_I. Ultimately, the y axis is displayed as the square root of the bin counts to emphasize spikes with low occurrence but potentially large amplitudes (Brock 1986).

The advantages of median filtering are its robustness, simplicity, and that its effectiveness is not altered by spikes that are frequent or high in amplitude. Its disadvantage is that it cannot detect spikes of more than a prescribed number of consecutive points N, which is a function of the window size, $2N + 1$ (Brock 1986). Furthermore, experimentation suggests that this procedure will detect some spikes even in clean data. Therefore, some preliminary qualitative examination of a few random samples of a given dataset should be performed to broadly assess its quality, before automating a despiking algorithm over the entire set.

For this work we adopt the median filter algorithm as described by Brock (1986, hereafter BR). We use an order $N = 3$ (i.e., a window of seven points), since such a filter will be sensitive to spikes of up to three consecutive points (Brock 1986), which by definition is the maximum number of successive points within a spike (VM; Mauder and Foken 2011). The raw data are first normalized (by subtracting its mean and dividing by the standard deviation) before applying the BR method. This ensures that the central population for the histogram of D_I is centered near zero. Then, we locate zero bin sizes on either side of the central population (R on the positive side; L on the negative side). Next, points in the original signal for which $D_I > R$ and $D_I < L$ are flagged as spikes.

c. *Phase-space thresholding*

Another despiking algorithm was derived by Goring and Nikora (2002) for cleaning measurements made using an acoustic Doppler velocimeter. The scheme is

known as phase-space thresholding. The procedure requires that three axes be derived: the signal itself U (x axis); its first surrogate time derivative, approximated as a discrete difference over a constant sampling period, ΔU (y axis); and its second surrogate time derivative, also approximated discretely, $\Delta^2 U$ (z axis). These axes define three planes (x–y, x–z, and z–y) whose superposition generates a three-dimensional cluster of points within which most of the data lie. Spikes are those points that reside a certain distance from the main cluster. That distance is based on the universal threshold, which is the expected absolute maximum value of a normally distributed, standard, random series of independent values (Goring and Nikora 2002), the mathematics of which are explained in detail by Donoho and Johnstone (1994). This threshold, as used by Goring and Nikora (2002), is calculated independently for the x, y, and z axes:

$$\lambda_U \sigma = [2 \ln(n) \sigma]^{1/2}, \quad (1)$$

where n is the sample size, and σ is the standard deviation of U, ΔU, and $\Delta^2 U$ (Goring and Nikora 2002). By evaluating $\lambda_U \sigma$ for each axis, a three-dimensional bounding ellipsoid is generated within which lie acceptable values, and outside of which are spikes. As spikes are detected, the algorithm replaces them and runs iteratively until no more are detected. This method has the advantage that it does not require the calibration of independent parameters for streamflow measurements (i.e., $\lambda_U \sigma$ is a function of the data itself), and that it can detect multiple spike events. Some caveats are that for $\lambda_U \sigma$ to be valid, the data need to have its mean removed and be detrended. In addition, since the universal threshold is based on "normal probability distribution theory" (Goring and Nikora 2002, p. 118), the data should ideally be independent and normally distributed. Certain datasets, especially those with coherent turbulence, may violate these criteria enough that the universal threshold is less amenable. Finally, this method also requires iterations, owing to the fact that spike replacement reduces the value of σ (Goring and Nikora 2002).

In this study, we adapt the method of Goring and Nikora (2002, hereafter GN). In the GN method, the raw signal is first detrended and its mean is removed prior to processing. Ten iterations are allowed before outputting the results, and spikes are removed with each subsequent iteration performed on the remaining values. In our initial testing of this algorithm on samples of real data, we found some oversensitivity of the universal threshold to the rapid temperature changes associated with the ramp shapes resulting from coherent structures. Therefore, in GN, the universal threshold is increased by an amount determined via empirical calibration.

d. *Spectral analysis by wavelet transform*

In their paper, GN also describe a method for spike detection called wavelet thresholding. Wavelet thresholding is inspired by the denoising technique called wavelet shrinkage, proposed by Donoho and Johnstone (1994). Wavelet shrinkage takes advantage of the compact property of the wavelet kernel, which allows signal behaviors to be located in time (Donoho and Johnstone 1994, and citations therein). Since the primary signal can be described by a few wavelet coefficients, other coefficients can be discarded as noise (i.e., "shrunk," or set to zero) if they fall below the threshold defined by Donoho and Johnstone (1994).

According to GN, wavelet thresholding is similar except that one seeks to remove spikes above a threshold versus shrinkage, where the desire is to remove noise below a threshold. Wavelet thresholding is performed by evaluating the first (lowest scale) wavelet coefficient where the mother wavelet has unit dilation. Points where this coefficient exceeds the universal threshold are spikes. Caveats are that the signal must have its mean removed and be zero padded, and that the process is iterated with spike replacement after each round. Also, missing values pose a problem for this methodology, so point replacement is necessary.

In the current study, we adapt a wavelet thresholding method (WT). The raw signal is first normalized to a zero mean and ± 1 standard deviation. Next, the normalized signal is decomposed using the Daubechies mother wavelet (Daubechies 1992). From there, the low-frequency portion of the signal obtained from that decomposition is used to reconstruct a low-pass-filtered signal, also known as an approximate signal. The distance (D_I) of the points in the normalized signal from the low-frequency reconstructed signal is then evaluated. Then, as with the BR method, a histogram of differences (D_I) is produced. This is done rather than using a universal threshold because, as stated previously, this threshold has been found to require additional calibration for the signals used in micrometeorology.

e. *Quantifying overall performance of the algorithms*

In addition to evaluating the ability of each algorithm to detect known spike points in the data, we also compare their performances via a metric that is based on the *Euclidean* distance, or the root-mean-square deviation (RMSD). We calculate the reduction in RMSD relative to a reference signal as

$$\mathrm{RMSD}_r = -100\left[1 - \left(\frac{\mathrm{RMSD}_a}{\mathrm{RMSD}_w}\right)\right], \quad \text{with;} \quad (2)$$

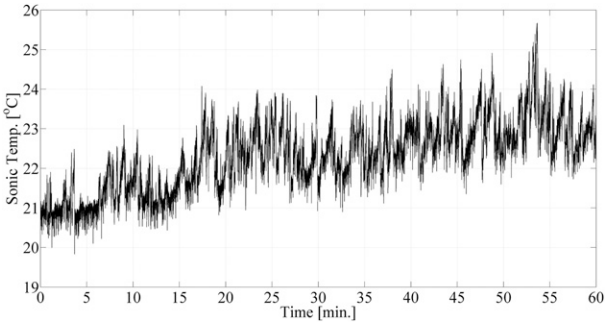

FIG. 1. A 60-min sample of actual 20-Hz sonic temperature data for a summer case in an Alaskan black spruce boreal forest. The data were taken 12 m above ground over a canopy that has an average height of 4.7 m.

$$\mathrm{RMSD}_a = \left\{ \frac{1}{N} \sum_{1}^{N} [\mathrm{despike}(X_s) - X_r]^2 \right\}^{1/2}, \quad \text{and;} \tag{3}$$

$$\mathrm{RMSD}_w = \left[\frac{1}{N} \sum_{1}^{N} (X_s - X_r)^2 \right]^{1/2}. \tag{4}$$

Here, X_s is the signal of interest to be treated with the despike(X) algorithm $\in$ {BR, GN, WT, VM}. Whenever a spike is identified, a missing value is imputed and subsequently linearly interpolated. The reference signal X_r is created by imputing missing values for all known spike points in X_s and subsequently linearly interpolated. Thus, X_r represents the case that the despike(X) algorithm was able to remove all true spikes, without false detections. Consequently, RMSD_r will indicate a reduction of -100% if all true but no falsely detected spikes are removed, and $\mathrm{RMSD}_r > -100\%$ if only part of all true spikes and/or falsely detected spikes is removed.

In short, the larger the reduction in RMSD_r, the better the algorithm cleaned the signal. As the RMSD itself is based on Gaussian statistics, RMSD_r thus reflects the ability of the despike(X) algorithm to rectify the standard deviation of the signal of interest.

3. Data

a. Synthetic signal

To perform this analysis, a 1-h synthetic temperature signal is generated at a 20-Hz sampling frequency (i.e., 72 000 points) based on Højstrup (1993):

$$S_j = (S_{j-1}R) + (1 - R)\mu + \sigma[(1 - R^2)]^{1/2} G(\mu, \sigma). \tag{5}$$

Here, a point S_j is a function of the previous point S_{j-1} and R (the point-to-point correlation); μ is the signal mean and σ is its standard deviation, while G is a random number based on the parameters μ and σ (it should be noted that the resultant signal S will have an actual mean higher than the prescribed μ; this is not of concern here, since only the local variation is important for despiking). Appropriate values for R and σ were selected empirically by comparison with a real 20-Hz sonic temperature series taken from the boreal forest of interior Alaska during summertime (Fig. 1). Since temperature ramps occur in scalar time series within and over canopies (Gao et al. 1989; Starkenburg et al. 2013), while inverse ramps, or canopy waves, occur during more stable conditions (Lee et al. 1997), we introduce a series of forward and inverse ramp shapes whose amplitude and duration are based on our experience working with such data (Starkenburg et al. 2013). This signal is hereafter the clean synthetic signal (CSS; Fig. 2). Afterward, a spiked version of the synthetic signal (SSS) was generated by inserting 720 spiked points (1% of the data series). The spike locations were randomly selected, and divided into 360 single spikes, 240 spikes as double (consecutive) events, and 120 spikes as triple (consecutive) events. Each spike (single, pair, or triplet) is shifted by $+1.5°$ of temperature if its first position location is an even number, and by $-1.5°$ if its location is odd (Fig. 2). This spike amplitude in SSS (1.5°) is twice the standard deviation of the CSS ($\pm 0.75°$). The distribution of D_I based on the median filter method (BR) is unimodal for the CSS, and multimodal for the SSS (Fig. 3).

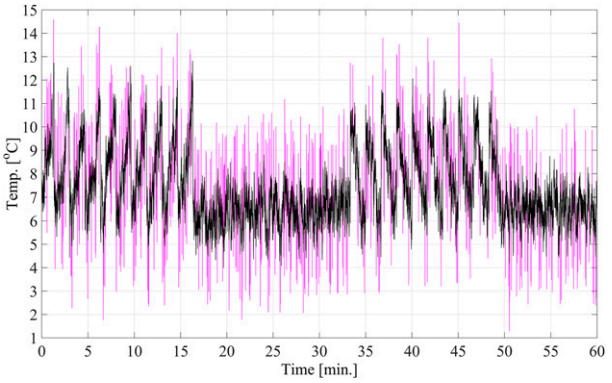

FIG. 2. SSS in magenta, with the CSS superimposed in black for comparison. This signal emulates a 60-min turbulent temperature at a 20-Hz sampling rate. Units are scaled to 1°C increments.

b. Real signal

A very clean signal or an excessively spiked signal does not provide a meaningful evaluation of a despiking methodology. In this study, we use a 20-Hz, 6-h time series (0800–1400 UTC 21 October 2008) of CO_2 measured 50 m above ground at a micrometeorological tower on a flat field site in Lindenberg, Germany (Metzger

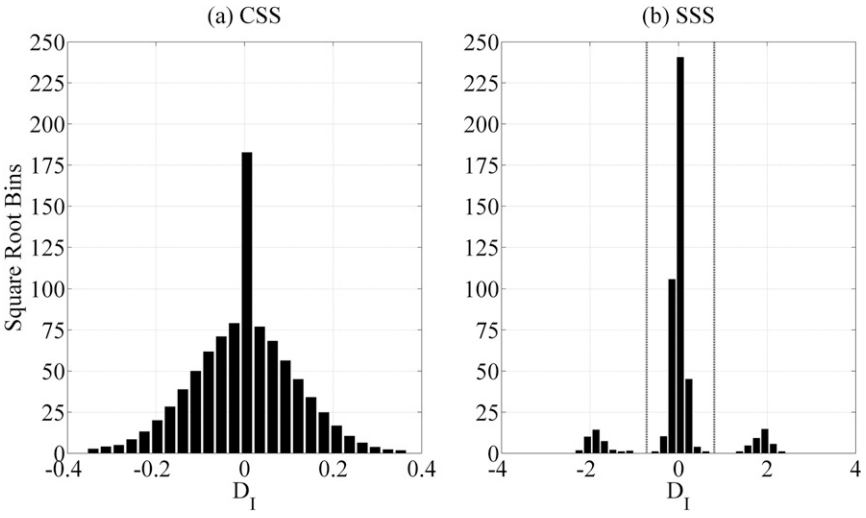

FIG. 3. Histogram of the distance between the normalized raw signal and the median filter D_I for (a) CSS and (b) SSS. Vertical lines in (b) indicate the first zero bin values to the left and right of 0, between which lie the quality data and beyond which points are considered spikes. Median filter value $N = 3$ and the window size $= 7$; histogram bin size for D_I is 25.

et al. 2012). This signal has two qualities that make it amenable to this evaluation. First, the CO_2 measurement contains severe disruptions due to signal communication (data downlink) every hour that produce obvious spikes of unreasonable amplitude (a standard flux processing algorithm would reject such extreme outliers during an initial range test). The signal also contains several smaller (but still obvious) spikes between these systematic outliers that can be qualitatively discerned. The extreme spikes at each hour can be removed, thus providing a reduced version of the signal for a second round of processing (Fig. 4).

The discussion that follows is the result of testing the despiking algorithms on turbulent temperature and CO_2 data. The characteristics of other scalar time series, such as that for CH_4, may require additional consideration owing to their unique behaviors and spatial distributions (e.g., Felber et al. 2015).

4. Results and discussion

a. *Despiking synthetic data*

GN state that the universal threshold is amenable to their analysis because "the velocity spectrum at low frequencies approximates the white-noise spectrum" (Nikora and Goring 2000, p. 124), and that their data are approximately normal, such that the standard deviation sufficiently describes the distribution. Qualitative evaluation on real data reveals that the sensitivity of GN to second derivatives sometimes causes false detections in micrometeorological data in the presence of coherent turbulence where ramp shapes do not preserve the second-order momentum in the data distribution and therefore violate the normal distribution required for the universal threshold to perform. Thus, we perform a sensitivity test on GN using CSS and SSS, in which we divide the universal threshold by varying coefficients. Ultimately, a coefficient of 0.48 was selected as a trade-off between correct detection and overdetection.

Next, all algorithms are run on the CSS to determine how they treat clean data (Table 2). For the CSS, GN detects over 500 spikes, many associated with ramp features. WT and BR detect 11 and 10 spikes, respectively, while VM detects no spikes. The bin size in Table 2 indicates that the histogram of D_I required a resolution of 103 bins (i.e., 25 were first tried, then 51, and then 103) to find zero bin counts.

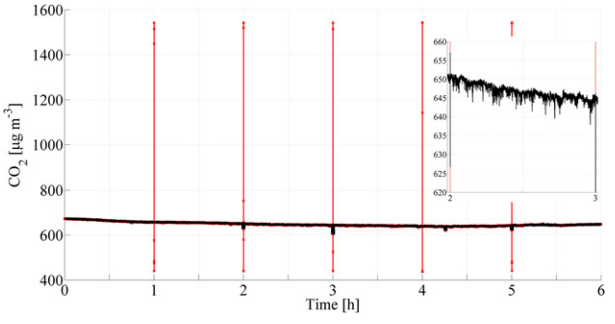

FIG. 4. The real CO_2 data series showing obvious outlier points every hour, and the reduced signal when outliers below 600 and above 750 $\mu g\,m^{-3}$ are replaced (black). A portion of the reduced signal in the interval (2–3 h) is shown in the insert.

TABLE 2. Performance of despiking methods for the CSS. First column is the despiking method, second column is the total number of spikes detected, third column is the threshold used (if applicable), and fourth column is the bin size used to isolate zero bins in the distribution of D_I (if applicable).

Method	Total	Threshold	Bin size
GN	525	0.48	—
WT	11	—	103
BR	10	—	103
VM	0	—	—

The algorithms are then applied to the SSS. For each method, the distance of all detected spikes in the SSS from their corresponding point in the CSS was evaluated (D_X). For GN, a large amount of false detections ($D_X = 0$) occurs, as well as a large amount of true spikes ($D_X = \pm 1.5$); many of the false detections are near the sharp drop in ramp shapes associated with coherent structures (most of the points in the ramp remain; so that in practice, we do not expect significant problems with flux calculations on data despiked with GN) (Fig. 5). Results for WT are an improvement, as it can be seen that most of the detected spikes are correct ($D_X = \pm 1.5$). The overdetections from WT are also often near ramp features; however, the sensitivity to ramps is less than with GN. The results for BR show that all but three of the detections are correct, while the results of VM reveal that all detected spikes are correct.

Based on these results, it might appear that VM is the optimal scheme because it detected no spikes in the CSS and had only correct detections in the SSS. However, Table 3 reveals the total detections, correct detections, missed detections, and overdetections in the SSS for each algorithm. VM detects only 167 of the 720 spiked points in the SSS (553 spikes were missed). In the SSS, the modest amplitude of the spikes may be problematic for a pure statistically based method that relies on the mean (VM) to discern outliers. Further, it is noted that while GN and WT have some sensitivity to ramps, most of the spikes found by VM are not within ramp shapes (we noted similar results with actual data). Since ramps are local changes in the signal mean, it may be difficult for an algorithm such as VM to locate spikes within ramps.

Table 3 also shows that for GN, 705 of the detected spikes were correct. However, this came at a great cost of overdetection (297 false detections). By comparison, WT locates 757 total spikes, 711 of which are correct (with 46 overdetections). BR detected 718 spikes, 715 of which are correct (only 3 overdetections). Thus, BR appears to perform optimally. The bin size in Table 3 shows that the initial bin size of 25 was sufficient to isolate spikes for BR, but 207 bins were needed in WT. The distribution of D_I from WT is slightly smoother than

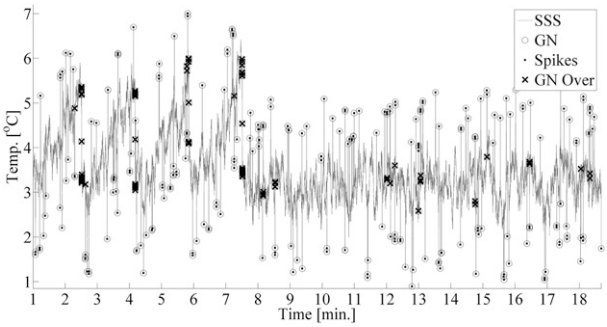

FIG. 5. Testing GN on the SSS. Circles (GN) are spikes detected by GN, black dots are known spike positions, and X marks indicate overdetections by GN. Units are scaled to 1°C increment.

that from BR (Fig. 6). The density of the histogram from WT in Fig. 6b is partially the result of the greater bin size. However, even when the bin size for WT is held at 25 (Fig. 6c), the histogram of D_I for the WT is still smoother than for the BR method. This is likely due to the stepwise nature of the windowed median filter in BR as compared to the smoother reconstructed signal from WT. Ultimately, in terms of correct detections, both WT and BR perform about equally well. However, BR has the fewest false detections, making it the preferred method for this analysis.

b. *Despiking real data*

Since GN requires extensive calibration for use on micrometeorological data, we chose to test the real data series using only WT, BR, and VM. This is done in two steps: 1) apply the algorithms to the raw CO_2 signal, including the extreme outliers that occur each hour; and then 2) remove and replace those outliers to provide a reduced version of the CO_2 data, and then run the algorithms again. The test assesses the ability of the algorithms over a range of spike amplitudes. The point replacement simply prevents the problem of missing values that bias the spectral methodology of WT. Since

TABLE 3. Performance of despiking methods for the SSS. The first column is the despiking method, the second column is the total number of spikes detected, the third column is the number of correct detections, the fourth column is the number of spikes missed by the algorithm, the fifth column is the number of incorrect (over) detections, the sixth column is the threshold used (if applicable), and the seventh column is the bin size used to isolate zero bins in the distribution of D_I (if applicable).

Method	Total	Correct	Missed	Over	Threshold	Bin size
GN	1002	705	15	297	0.48	—
WT	757	711	9	46	—	207
BR	718	715	5	3	—	25
VM	167	167	553	0	—	—

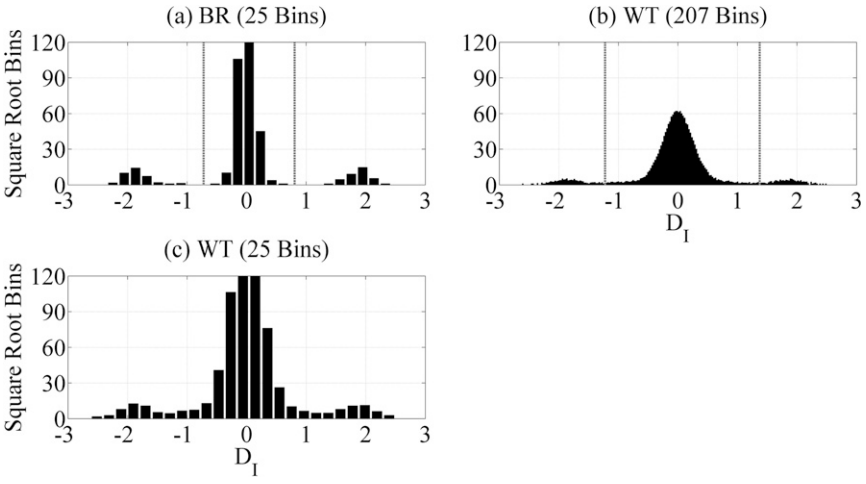

FIG. 6. (a) Distribution of the normalized raw SSS minus the median filter for BR, and (b) as in (a), but minus the reconstructed signal for WT. (c) As in (b), but with only 25 bins for better comparison to (a). Vertical lines indicate the first zero bin values on either side of zero. Note that the *y* axis is terminated at 120 for visual clarity.

replacing points (and the method for how to do so) is neither a goal nor necessarily a recommendation of this work, we use a simple procedure that replaces the extreme spikes with a value that is the mean of the 20 surrounding points, simply to reduce their amplitudes (not to accurately replace them for use in any analysis).

Table 4 provides the results of step 1, where the raw CO_2 signal is despiked. In this step, a broad-brush approach was taken to flag the most extreme values, as defined qualitatively by being less than 600 or greater than $750\,\mu g\,m^{-3}$. The positions of these 39 instances were retained as known spike points. WT and BR found a similar number of spikes (38 and 37, respectively), all of which were contained in the 39 extreme values originally flagged. Table 4 shows that both WT and BR required a bin size of 51 (meaning two iterations on the bin size) in order to find zero bins, while Fig. 7 shows the resulting distributions of D_I for both methods. The distribution of D_I has many zero bin values, as expected when the spiked points have such extreme amplitudes. However, the distribution for WT is again slightly smoother.

Also in Table 4 are the results of VM, who detected 557 spikes (only 6 of which were the ones flagged as extreme outliers). The outliers have severe amplitudes and occur as multiple spiked points every hour. Therefore, it is possible that the mean is biased by the magnitude of these points. It is also possible that the spikes go undetected because they are more than three consecutive points. Regarding the latter, it should be noted that these extreme spikes oscillate between being higher or lower than the local mean, such that no more than three consecutive points lie well above, before the next point lies well below, the local mean. This appears to be important, because BR was able to detect multiple extreme outlier points (e.g., 7 in a row), despite that it had an order $N = 3$ (window size of $2N + 1 = 7$), which limits it to the detection of three or fewer consecutive points. Ultimately, VM is unable to detect most of the extreme values, while BR and WT had more success.

The next step is to compare the performances of WT, BR, and VM on the reduced CO_2 signal. In this instance, the 39 most extreme outliers (< 600 or $> 750\,\mu g\,m^{-3}$) from the raw signal are replaced and reduced successfully enough that none of the despiking algorithms detected these filled points. However, five spikes that result each hour from the signal communication were not removed because their values were not extreme (< 600 or $> 750\,\mu g\,m^{-3}$) and therefore still remain in the reduced signal. Furthermore, seven other points that are not associated with the systematic hourly interference also exist, and can now be visually selected and included as spikes. Together, these 12 new points are the ''known'' spikes for the reduced signal.

After running the algorithms on the reduced signal, we note that VM located 565 spikes, 11 of which are correct detections but 554 of which are overdetections (Table 5; Fig. 8). This means that in the reduced signal,

TABLE 4. As in Table 3, but for the performance of WT, BR, and VM on the raw CO_2 signal containing extreme outliers.

Method	Total	Correct	Missed	Over	Bin size
WT	38	38	1	0	51
BR	37	37	2	0	51
VM	557	6	33	551	—

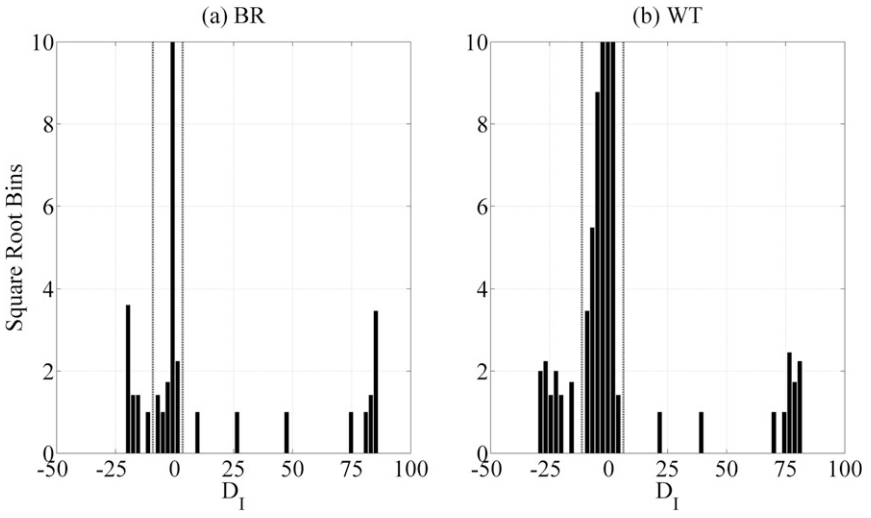

FIG. 7. (a) Distribution of the normalized raw CO_2 signal minus the median filter for BR, and (b) as in (a), but minus the reconstructed signal for WT. Vertical lines indicate the first zero bin values on either side of zero. Note that the *y* axis is terminated at 10 for visual clarity.

the spike amplitude is not too extreme for VM to be able to correctly detect them. However, despite that VM can detect most of the known spikes, it also detects a large number of points that visually do not appear to be spikes. These overdetections occur during the intrusion of CO_2-depleted eddies. These periods are composed of sufficient data points (order 100) such that they should not be considered outliers. Furthermore, during those periods, the points VM flags as spikes are not necessarily the local extrema. When comparing the results across methods, it becomes apparent that those based on Gaussian statistics are more susceptible to producing artifacts, and that their sensitivity to artifacts is modulated by parameters such as window size and threshold.

By comparison, WT correctly detects 10 spikes (missing 2), while BR correctly detects 8 spikes (missing 4). WT and BR both still require a bin size of 51 to isolate zero bin values within the distribution of D_I. Furthermore, the distribution of D_I is similar for BR and WT, and the range of accepted data also matches closely (Fig. 9). Neither WT nor BR overdetected. Based on this result, WT performed slightly better on the reduced signal. This raises the question as to why BR and WT perform similarly on this real dataset, while BR was superior when applied to the SSS. One difference is that the spike amplitude in the real signal is larger than in the SSS, even for the reduced version. Also, ramp shapes are less sharp in the real signal owing to an expected amount of instrument noise, which softens their edges. It may be that sharp, systematic ramps and low spike amplitudes in the SSS hinder the performance of WT (and GN as well).

In Table 6, we tabulate the reduction in RMSD for each despiking algorithm relative to a reference signal (section 2e). We also include additional algorithms from Table 1 as a reference for comparison (italics in Table 6). For statistical time window–based algorithms that use the mean, successful despiking highly depends on the statistical measures used and specific parameterizations. For these algorithms (VM; Mauder and Foken 2011; Fratini and Mauder 2014), we obtain comparatively small and varying reductions of $\mathrm{RMSD}_r \geq -36\% \pm 51\%$. By comparison, statistical median-based windowed algorithms are more successful (Metzger et al. 2012; Mauder et al. 2013) but still largely variable across datasets with an RMSD_r of $-59 \pm 52\%$ and $-58 \pm 52\%$, respectively. The optimal algorithms are BR and WT, which consistently have large RMSD_r values across all data. BR is excellent, with $\mathrm{RMSD}_r = -92 \pm 6\%$. While WT performs even better in raw numbers ($\mathrm{RMSD}_r = -93 \pm 3\%$), it is most optimal for data not dominated by sharp ramp shapes, and especially for cases where there are no missing data values. We caution that even optimal procedures such as BR and WT will detect spikes even in clean data, and thus we recommend some qualitative review of a few random data samples from each download to get a general sense of data quality before despiking a large amount of data.

For practical implementation, despiking performance is important, but so is the ability to automate the process

TABLE 5. As in Table 4, but for the performance of WT, BR, and VM on the reduced CO_2 signal.

Method	Total	Correct	Missed	Over	Bin size
WT	10	10	2	0	51
BR	8	8	4	0	51
VM	565	11	1	554	—

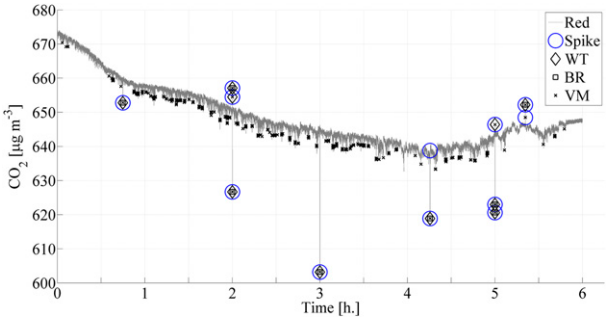

FIG. 8. Result of running WT, BR, and VM on the reduced CO_2 signal (after removing and gap filling the most extreme outlier points as described in the text). Gray is the reduced signal, and "spike" indicates known spike positions (circles). WT, BR, and VM are where the respective algorithms detected spike points.

and the overall computational efficiency. From a despiking perspective, BR performs very well, followed by WT and median-based statistical methods (Metzger et al. 2012; Mauder et al. 2013). The mean-based statistical methods (VM; Mauder and Foken 2011; Fratini and Mauder 2014) and GN trail behind, which is also the reason for Fratini and Mauder (2014) to suggest upgrading EddyPro (LI-COR 2012) from the mean-based method to the median-based method. Regarding automation for use with micrometeorological data, BR and the median- and mean-based statistical methods rank highest, with WT next in line (owing to the dependence of WT on gap-free data), followed by GN (owing to its calibration required to adjust the universal threshold for coherent turbulence). Computationally, the most efficient was WT, followed by BR and GN, and trailed by the median- and mean-based statistical methods, each group separated by approximately one order of magnitude runtime. Equal emphasis on these attributes would place BR ahead of WT, followed by median- and mean-based statistical methods, and trailed by GN. Ultimately, the choice of the most appropriate method will depend on the emphasis that a specific use puts on the different attributes.

5. Conclusions

In this analysis, different methodologies for despiking high-frequency resolution data series were reviewed, and four of them were adapted so that their performance could be evaluated using both a synthetic signal and a real signal. Specifically, we evaluate a statistical method based on VM, a phase-space thresholding scheme based on GN, a nonlinear median filter with threshold logic based on BR, and a wavelet thresholding technique inspired from Donoho and Johnstone (1994) and GN called WT. We acknowledge that outside of obviously unrealistic values, visual spike confirmation is always somewhat subjective, and therefore the evaluation of the performance of these despiking algorithms are subject to those limitations.

Across all the signals tested, WT and BR have the largest and most stable reduction in root-mean-square deviation compared to a reference value, suggesting optimal despiking performances. This also holds true in comparison to widely used time window–based despiking methods, which on average perform 34% and 71% worse

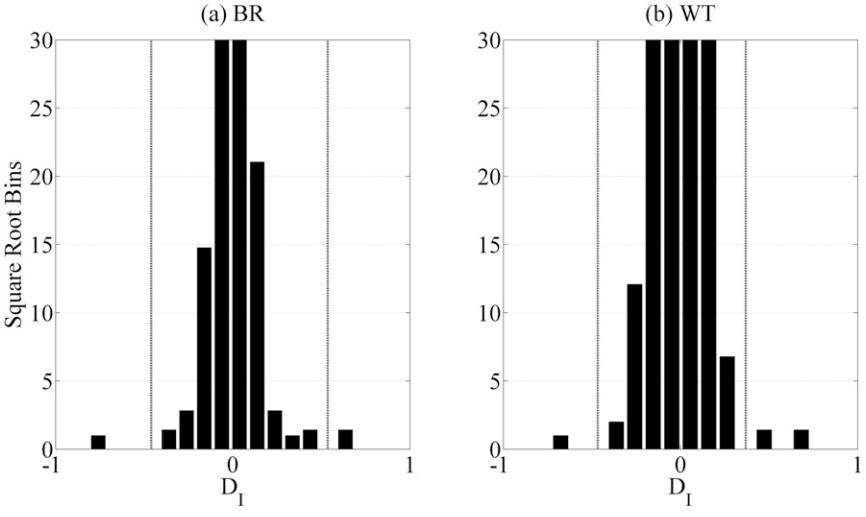

FIG. 9. (a) Distribution of the normalized reduced CO_2 signal minus the median filter for BR, and (b) as in (a), but minus the reconstructed signal for WT. Vertical lines indicate the first zero bin values on either side of zero. Note that the *y* axis is terminated at 30 and that the *x* axis is set between −1 and +1 for visual clarity.

TABLE 6. Relative reduction in $RMSD_r$ due to despiking. The first column is the despiking method, including several widely used procedures listed in Table 1 for context (italics). The second through fourth columns are the $RMSD_r$ for the SSS, raw, and reduced CO_2 signals, respectively. The fifth and sixth columns are the $RMSD_r$ mean and standard deviation for each algorithm across all signals, respectively.

Method	SSS	CO_2 raw	CO_2 reduced	Mean
WT	−90%	−97%	−92%	−93% ± 3%
BR	−94%	−97%	−85%	−92% ± 6%
Metzger et al. (2012)	*−0%*	*−100%*	*−77%*	*−59% ± 52%*
Mauder et al. (2013)	*−0%*	*−100%*	*−74%*	*−58% ± 52%*
VM	−12%	−1%	−94%	−36% ± 51%
Fratini and Mauder (2014)	*−17%*	*−1%*	*−64%*	*−28% ± 33%*
Mauder and Foken (2011)	*−0%*	*−0%*	*−0%*	*−0% ± 0%*
GN	−87%	—	—	—

for implementations using median- and mean-based statistical measures, respectively. GN often overdetects in the presence of sharp ramps (coherent turbulence) in the synthetic signal. On the other hand, VM underdetects in the synthetic signal with sharp ramps and small amplitude spikes, while it overdetects in the real signals with larger amplitude spikes. BR and WT both do a superior job of detecting most of the known spikes with minimal overdetection.

We conclude that WT and BR are both adequate methods. We recommend BR as the most versatile choice, however, because it does not appear sensitive to low-frequency turbulence (i.e., ramp shapes), which are prevalent in many micrometeorological data taken over plant or forest canopies. Also, BR is able to handle missing values, even one that will strongly skew the results of WT. Therefore, use of BR does not require point replacement, which is an advantage since point replacement is not always desirable depending on the data type and its application. Moreover, BR is highly adaptive with a minimal set of parameters and is computationally efficient. Altogether, this makes BR most suitable also for large-scale, automated, and standardized processing in surface–atmosphere exchange networks, such as the Integrated Carbon Observation System (ICOS) and the National Ecological Observatory Network (NEON). One of the most important results of this study is that a thorough testing of despiking algorithms over different signals and using different metrics is required to fully determine their robustness. To this end, these despiking algorithms should be tested with many datasets over a variety of physical and meteorological conditions to determine long-term performance.

Acknowledgments. This research was supported by the Alaska NASA EPSCoR Program Award NNX10NO2A, by the Alaska Space Grant Program, and by the "New GK-12 Program: The CASE (Changing Alaska Science Education) for Enhancing Understanding of Climate Change" NSF (DGE-0948029). Data from the Falkenberg tower were kindly provided by Deutscher Wetterdienst Meteorologisches Observatorium Lindenberg-Richard-Aßmann-Observatorium. We thank J. P. Leps and K. Jantze for the CO_2 datasets. The National Ecological Observatory Network is a project solely sponsored by the National Science Foundation and is managed under cooperative agreement by NEON, Inc. This material is based upon work supported by the National Science Foundation under Grant DBI-0752017. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Datasets for this paper can be requested from Professor G.J. Fochesatto (gjfochesatto@alaska.edu), Geophysical Institute, University of Alaska Fairbanks.

REFERENCES

Beaton, A. E., and J. W. Tukey, 1974: The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, **16**, 147–185, doi:10.1080/00401706.1974.10489171.

Brock, F. V., 1986: A nonlinear filter to remove impulse noise from meteorological data. *J. Atmos. Oceanic Technol.*, **3**, 51–58, doi:10.1175/1520-0426(1986)003<0051:ANFTRI>2.0.CO;2.

Daubechies, I., 1992: *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 61, Society for Industrial and Applied Mathematics, 350 pp.

Donoho, D. L., and I. M. Johnstone, 1994: Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455, doi:10.1093/biomet/81.3.425.

Felber, R., A. Münger, A. Neftel, and C. Ammann, 2015: Eddy covariance methane flux measurements over a grazed pasture: Effect of cows as moving point sources. *Biogeosciences*, **12**, 3925–3940, doi:10.5194/bg-12-3925-2015.

Fochesatto, G. J., and J. Sloan, 2008: Signal processing of multicomponent Raman spectra of particulate matter. *Int. J. High Speed Electron. Syst.*, **18**, 277–294, doi:10.1142/S0129156408005345.

Foken, T., and B. Wichura, 1996: Tools for quality assessment of surface-based flux measurements. *Agric. For. Meteor.*, **78**, 83–105, doi:10.1016/0168-1923(95)02248-1.

——, and Coauthors, 2010: Energy balance closure for the LITFASS-2003 experiment. *Theor. Appl. Climatol.*, **101**, 149–160, doi:10.1007/s00704-009-0216-8.

Fratini, G., and M. Mauder, 2014: Towards a consistent eddy-covariance processing: An intercomparison of EddyPro and TK3. *Atmos. Meas. Tech.*, **7**, 2273–2281, doi:10.5194/amt-7-2273-2014.

Gallagher, N. C., and G. L. Wise, 1981: A theoretical analysis of the properties of median filters. *IEEE Trans. Acoust. Speech Signal Process.*, **29**, 1136–1141, doi:10.1109/TASSP.1981.1163708.

Gao, W., R. H. Shaw, and K. T. Paw U, 1989: Observation of organized structure in turbulent flow within and above a forest canopy. *Bound.-Layer Meteor.*, **47**, 349–377, doi:10.1007/BF00122339.

Goring, D. G., and V. I. Nikora, 2002: Despiking acoustic Doppler velocimeter data. *J. Hydraul. Eng.*, **128**, 117–126, doi:10.1061/(ASCE)0733-9429(2002)128:1(117).

Hill, W., and D. Rogalia, 1992: Spike-correction of weak signals from charge-coupled devices and its applications to Raman spectroscopy. *Anal. Chem.*, **64**, 2575–2579, doi:10.1021/ac00045a019.

Højstrup, J., 1993: A statistical data screening procedure. *Meas. Sci. Technol.*, **4**, 153–157, doi:10.1088/0957-0233/4/2/003.

Lee, X., H. H. Neumann, G. Hartog, R. E. Mickle, J. D. Fuentes, T. A. Black, P. C. Yang, and P. D. Blanken, 1997: Observation of gravity waves in a boreal forest. *Bound.-Layer Meteor.*, **84**, 383–398, doi:10.1023/A:1000454030493.

——, W. J. Massman, and B. E. Law, Eds., 2004: *Handbook of Micrometeorology: A Guide for Surface Flux Measurement and Analysis*. Atmospheric and Oceanographic Sciences Library, Vol. 29, Springer, 250 pp.

LI-COR, 2012: EddyPro 3.0 help and user's guide: Despiking and raw data statistical screening. Accessed 8 June 2015. [Available online at http://envsupport.licor.com/envhelp/EddyPro3/Content/Topics/Despiking_Raw_Stat_Screening.htm.]

Mauder, M., and T. Foken, 2006: Impact of post-field data processing on eddy covariance flux estimates and energy balance closure. *Meteor. Z.*, **15**, 597–609, doi:10.1127/0941-2948/2006/0167.

——, and ——, 2011: Documentation and instruction manual of the eddy-covariance software package TK3. Universität Bayreuth Dept. of Micrometeorology Work Results 46, 60 pp.

——, M. Cuntz, C. Drüe, A. Graf, C. Rebmann, H. P. Schmid, M. Schmidt, and R. Steinbrecher, 2013: A strategy for quality and uncertainty assessment of long-term eddy-covariance measurements. *Agric. For. Meteor.*, **169**, 122–135, doi:10.1016/j.agrformet.2012.09.006.

Metzger, S., W. Junkermann, M. Mauder, F. Beyrich, K. Butterbach-Bahl, H. P. Schmid, and T. Foken, 2012: Eddy-covariance flux measurements with a weight-shift microlight aircraft. *Atmos. Meas. Tech.*, **5**, 1699–1717, doi:10.5194/amt-5-1699-2012.

Nikora, V. I., and D. G. Goring, 2000: Flow turbulence over fixed and weakly mobile gravel beds. *J. Hydraul. Eng.*, **126**, 679–690, doi:10.1061/(ASCE)0733-9429(2000)126:9(679).

Papale, D., and Coauthors, 2006: Towards a standardized processing of Net Ecosystem Exchange measured with eddy covariance technique: Algorithms and uncertainty estimation. *Biogeosciences*, **3**, 571–583, doi:10.5194/bg-3-571-2006.

Pasian, F., and A. Crise, 1984: Restoration of signals degraded by impulse noise by means of a low-distortion nonlinear filter. *Signal Process.*, **6**, 67–76, doi:10.1016/0165-1684(84)90052-5.

Phillips, G. R., and J. M. Harris, 1990: Polynomial filters for data sets with outlying or missing observations: Application to charge-coupled-device-detected Raman spectra contaminated by cosmic rays. *Anal. Chem.*, **62**, 2351–2357, doi:10.1021/ac00220a017.

Starkenburg, D., G. J. Fochesatto, A. Prakash, J. Cristóbal, R. Gens, and D. L. Kane, 2013: The role of coherent flow structures in the sensible heat fluxes of an Alaskan boreal forest. *J. Geophys. Res. Atmos.*, **118**, 8140–8155, doi:10.1002/jgrd.50625.

Vickers, D., and L. Mahrt, 1997: Quality control and flux sampling problems for tower and aircraft data. *J. Atmos. Oceanic Technol.*, **14**, 512–526, doi:10.1175/1520-0426(1997)014<0512:QCAFSP>2.0.CO;2.