Optimized Algorithm for Processing Outlier of Water Current Data Measured by Acoustic Doppler Velocimeter

Chunyi Zhong 1, Fang Yin 2, Junbo Zhang 1,3,4,5,6,*, Shuo Zhang 1,*, Rong Wan 1,4,* and Daisuke Kitazawa 7

1 College of Marine Sciences, Shanghai Ocean University, Shanghai 201306, China; chunyizhong@outlook.com
2 College of Ocean Science and Engineering, Shanghai Maritime University, Shanghai 201306, China; fangyin@shmtu.edu.cn
3 Key Laboratory of Sustainable Exploitation of Oceanic Fisheries Resources, Ministry of Education, Shanghai 201306, China
4 National Engineering Research Center for Oceanic Fisheries, Shanghai 201306, China
5 Key Laboratory of Oceanic Fisheries Exploration, Ministry of Agriculture and Rural Affairs, Shanghai 201306, China
6 Scientific Observing and Experimental Station of Oceanic Fishery Resources, Ministry of Agriculture and Rural Affairs, Shanghai 201306, China
7 Institute of Industrial Science, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba 277-8574, Japan; dkita@iis.u-tokyo.ac.jp
* Correspondence: zhangjunbo1985@gmail.com or jb_zhang@shou.edu.cn (J.Z.); s-zhang@shou.edu.cn (S.Z.); rwan@shou.edu.cn (R.W.)

Received: 25 July 2020; Accepted: 20 August 2020; Published: 25 August 2020

Abstract: In the process of pond culture, the usage of an aeration device can increase dissolved oxygen density and form a decent circulation which facilitates the collection of sludge. Acoustic Doppler Velocimeter (ADV) has been widely used to monitor the flow velocity, however, factors such as bubbles and suspended particles can affect the correlation coefficient and signal-to-noise ratio of ADV, which leads to the existence of outliers in velocity data. This study constructs the three-dimensional rousseeuw phase-space (3DRPS) method by optimizing the phase space threshold method and robust estimation method through two-step filtering and three-dimensional simultaneous measurement, where the outliers close to the real value can be detected more accurately and the iterative process can be reduced more effectively. The results show that the detection rate of the optimized 3DRPS method is approximately 99%. It is a promising method that effectively improves the accuracy of outlier detection and greatly reduces the phenomenon of over processing.

Keywords: acoustic doppler velocimeter; despiking; data processing; aquaculture pond

1. Introduction

With a rising global demand for food protein, the aquaculture of fish, crustaceans and mollusks has continually grown, faster than any other major food production sector, and significantly contributes to the increase in global aquatic production, i.e., 80 million tons in 2016 [1,2]. Due to the fast development of aquaculture, pond culture, a momentous intensive aquaculture production facility, has rapidly expanded and created enormous economic benefits; however, high organic sedimentation in the pond affects bacteria populations, which occasionally causes faster consumption of dissolved oxygen (DO) than its replenishment [3,4]. DO is a critical element of water quality in the aquaculture operation, and low DO concentration usually causes stress imposed on fish and even death due to suffocation.
Aeration machinery such as paddle wheel aerators, have been effectively applied to mitigate the risk of DO depletion in ponds. Aerators can facilitate the increase of the water flow velocity, which also transports the sludge created by hypoxic condition to the center of the pond for easy collection [5,6].

For the optimal arrangement of aerators in the pond, flow velocity monitoring is critical and Acoustic Doppler Velocimetry (ADV) is commonly used to fulfill this purpose. ADV is a velocity measuring device that is designed based on the principle of acoustic doppler effect. It uses ultrasonic technology to detect the velocity of water flow and executes the measurements of three-dimensional instantaneous velocity, turbulence and power spectral density [7,8]. Currently, ADV is widely utilized in field and laboratory hydrodynamic experiments, but high levels of noise and outliers have been reported in measurements using ADV [9,10]. Correlation coefficient (COR) and signal-to-noise ratio (SNR) are two important parameters with regard to ADV measurement accuracy, which can be easily affected by bubbles, large particle suspension, turbulence intensity and boundary layer. COR refers to the similarity between two pulse echoes, expressed as a percentage, where 100% denotes the consistency between transmission and reception of the echo, and 0% means that the two echoes are uncorrelated (the output results will be disturbed by noise). Typically, the ideal COR is between 70% and 100%. SNR is the ratio of signal response to background noise. Signal response is a measure of the intensity of the reflected acoustic signal recorded by the receiver of ADV, and it is primarily used to verify the sufficiency of particles in water. For the measurement of the average flow rate, the minimum requirement of signal to noise ratio is 5 dB; if the SNR is less than 5 dB, the ADV measurements will be negatively affected [11,12]. Some scholars have suggested that the velocity data with SNR less than 20 dB is not recommended [13].

In the aquaculture pond, the spray, bubble and resuspended particles that affect the COR and the SNR are mainly caused by the operation of aerators. A large number of outliers in monitored flow data will be easily generated [14–16]. In this regard, the postprocessing of doppler velocimetry data that can greatly improve the accuracy of velocity data for better analyzing the water velocity distribution of the pond and calculating the bottom friction velocity. This postprocessing technology of doppler velocimetry data is also quite useful for the analysis of current velocity in marine resource developments, such as the resource assessment of tidal current energy development.

At present, the main algorithms of data postprocessing are Robust Estimation (RE) Method (also called Rousseuw Method) and Phase Space Thresholding (PST) Method. The RE standardizes the sample data by the estimation of location value and the scale estimation value, then compares the set minimum passing value to diagnose the outliers. Construction of the detection threshold of PST is based on the mean value and standard deviation of the sample data. Also, PST keeps iterating while detecting outliers until the numbers of outliers are zero. For the velocity data obtained from complex environments, the detection accuracies of PST are higher than RE [11], nevertheless, numerous real data will be determined as outliers during each iteration. In addition, PST might be inoperative if over 6.7% outliers are found in the sample data [17]. According to the above reasons, an algorithm with high detection accuracy and less overtreatment that can effectively handle contaminated data under complex environmental conditions is highly required.

In this paper, a new method named three dimensional rousseuw phase-space (3DRPS) method was developed, which combines two-step filtering with three-dimensional simultaneous detection to void iteration, reduce the over processing rate and improve the detection accuracy. This method was applied to postprocess flow velocity data measured in aquaculture ponds, which could provide accurate velocity data of the velocity distribution and the estimation of bottom friction velocity for the follow-up study.
2. Materials and Methods

2.1. Data Postprocessing Algorithm

2.1.1. Robust Estimation

Since the mean value and the standard deviation are significantly affected by outliers, this approach is designed based on the median and the absolute deviations of sample, which was originally developed by Rousseeuw (1988) [18,19]. In this approach, real data will not be diagnosed as outliers, however, only the outliers with a large difference from the average could be detected. The algorithm of this approach is expressed as:

\[ M = \text{median}_{i=1,...,n}(u_i), \]  \hspace{1cm} (1)

\[ S = e(\text{median}_{i=1,...,n}|u_i - M|), \]  \hspace{1cm} (2)

\[ Z_i = (u_i - M)/S \]  \hspace{1cm} (3)

\[ O_i = |Z_i| - c \]  \hspace{1cm} (4)

where \( M \) is the robust location (m/s^{-1}), \( n \) is the number of the velocity data points, \( u_i \) is water current velocity data (m/s^{-1}), \( S \) is the scale estimator (m/s^{-1}), \( e \) is the estimator coefficient (1.483, –), \( Z_i \) is the standardized observations (–), \( O_i \) is outlier judgment (–), and \( c \) is the cutoff value (2.5, –). The outlier will be determined if the value of \( O_i \) is above 0. The deleted data points are usually replaced by the extrapolation from the two preceding points.

2.1.2. Phase Space Threshold Approach

As one of the most efficient algorithms for despiking the ADV velocity data, PST was originally developed by Goring and Nikora [14,20] and improved by Wahl [21]. The water current velocity \( u \) and its first and second derivatives can be plotted against each other by a Poincaré map to present an ellipsoid. The data located outside of the ellipsoid are detected as outliers, and the iteration process will be used to further exclude the outliers until the remaining points are all inside the ellipsoid. The replacement method for eliminated data points is cubic spline interpolation. Although the PST shows high detection accuracy, a large number of real data will be diagnosed as outliers and replaced during its iteration process. The algorithm can be described as:

\[ U = \text{mean}_{i=1,...,n}(u_i) \]  \hspace{1cm} (5)

\[ u'_i = u_i - U \]  \hspace{1cm} (6)

\[ \Delta u'_i = \Delta t \frac{du'_i}{dt} \]  \hspace{1cm} (7)

\[ \Delta^2 u'_i = (\Delta t)^2 \frac{d^2u'_i}{dt^2} \]  \hspace{1cm} (8)

\[ \sigma_{\Delta u'} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (\Delta u'_i - \Delta u')^2} \]  \hspace{1cm} (9)

\[ \sigma_{\Delta^2 u'} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (\Delta^2 u'_i - \Delta^2 u')^2} \]  \hspace{1cm} (10)

\[ \lambda = \sqrt{2 \ln n} \]  \hspace{1cm} (11)
where \( U \) is the average value of water current velocity (m/s\(^{-1} \)), \( u'_i \) is the fluctuated velocity (m/s\(^{-1} \)), \( \Delta u'_i \) is the first time derivative of velocity (m/s\(^{-1} \)), \( \Delta^2 u'_i \) is the second time derivative of velocity (m/s\(^{-2} \)), \( \sigma \) is the standard deviation of fluctuated velocity (\( \cdot \)), \( \lambda \) is the universal threshold (\( \cdot \)), and \( \theta \) is the rotation angle of the principal axis (rad). The major and minor axes of the ellipsoid are calculated by \( u'_i, \Delta u'_i, \) and \( \Delta^2 u'_i \), respectively.

2.1.3. Three Dimensional Rousseeuw Phase-Space Approach

The 3DRPS method is derived from RE and PST. Combined with the mobile window method [22], the 3DRPS utilizes a double detection to remove the outliers, which improves the detection accuracy and avoids the over processing phenomenon caused by the iteration (Figure 1). In the 3DRPS method, a window with fixed width (determined based on data sampling frequency) is considered. The RE and mobile window method are both used for despiking in the first-step filtration. RE detects outliers through Equations (1)–(4). After that, the robust location of each window is calculated according to Equation (13), in which the data located in the first and last windows will be dealt with using Equation (4). Then, the sample data are put into Equations (1)–(4) for secondary filtration through Poincare mapping, then the outliers close to real values can be identified and removed. Most of the ADV probes record three-dimensional velocity components (in some cases, two-dimensional) [21]. All velocity components recorded by ADV are orthogonal and interdependent. The outlier is usually one of the single dimension velocity components, however, in the process of converting the velocity component into orthogonal velocity, the affected velocity component will contaminate the rest of the orthogonal velocity components. The 3DRPS processes the water current velocity components of \( u, v \) and \( w \) simultaneously, which means if an outlier is detected in one of the three components, its corresponding data in the other two components will be eliminated to ensure the accuracy of the detected data. The eliminated data will be smoothed by cubic spline interpolation in the case of Runge phenomenon caused by polynomial interpolation.

\[
\theta = \begin{cases} \tan^{-1}\left(\frac{\sum_{i=1}^{n} u'_i \Delta^2 u'_i}{\sum_{i=1}^{n} u'_i^2}\right), & \Delta^2 u'_i \text{ versus } u'_i \\ 0, & \Delta u'_i \text{ versus } u'_i \\ 0, & \Delta^2 u'_i \text{ versus } \Delta u'_i \end{cases} \tag{12}
\]

\[
M_k = \text{median}_{j=1,...,w}(u_j), \quad j \leq \frac{w-1}{2} \tag{13}
\]
\[
M_k = \text{median}_{j=\frac{w}{2},...,j+\frac{w}{2}}(u_j), \quad \frac{w-1}{2} < j \leq n - \frac{w-1}{2} \tag{14}
\]
\[
M_k = \text{median}_{j=n-w,...,n}(u_j), \quad j > n - \frac{w-1}{2} \tag{15}
\]
\[
S_w = e(\text{median}_{j=1,...,w}|u_j - m_j|), \quad j \leq \frac{w-1}{2} \tag{16}
\]
\[
S_w = e(\text{median}_{j=\frac{w}{2},...,j+\frac{w}{2}}|u_j - m_j|), \quad \frac{w-1}{2} < j \leq n - \frac{w-1}{2} \tag{17}
\]
\[
S_w = e(\text{median}_{j=n-w,...,n}|u_j - m_j|), \quad j > n - \frac{w-1}{2} \tag{18}
\]

where \( k \) is the subscript for the sequence number of window in data set \( (-) \), \( n \) is the number of the velocity data points \( (-) \), \( w \) is the number of the total number of velocity data \( (-) \), \( W \) is the width of the window \( (-) \), \( j \) is the subscript for the sequence number of velocity data in each window \( (-) \), \( u_j \) is the water current...
velocity data (m·s\(^{-1}\)), \(M_k\) is the robust location in a mobile window (m·s\(^{-1}\)), and \(S_k\) is the scale estimator in a mobile window (m·s\(^{-1}\)).

**Figure 1.** Flowchart of data processing algorithm of three dimensional rousseeuw phase-space approach.

### 2.2. Data Sources

The data used in this paper include simulated data and measured data. Since the flow velocity data measured by ADV in the artificial ponds and natural lakes still contain a few outliers due to the influence of boundary layers or turbulence, the simulated data are used for algorithm verification to ensure zero outliers. The simulated data will be considered as clean data and then contaminated with some artificial outliers. The location and value of artificial outliers and the real detected outliers are recorded for the comparison and analysis of detection ability of 3DRPS. The measured data are used for application.

#### 2.2.1. Simulated Data

Clean velocity data are simulated by R software according to the flow velocity simulation formula and they are artificially contaminated with different degrees of outliers. Figure 2 is the schematic diagram of clean data and contaminated data of \(u\) direction with different degrees of outliers. The degrees of outliers (cases 1 to 7) are 1, 2, 5, 10, 15, 20 and 30%, respectively. The generation procedure of the contaminated data can be clarified as three steps: (A) randomly take a certain amount of value from 0 to 12,000 as the position of the outlier; (B) randomly generate a certain number of values in the range of −5 to 5 as outlier values; (C) use these outliers to replace the values in the clean data and then generate the contaminated data. Simulated sampling time is 120 s, and sampling frequency is 100 Hz. The simulation formula is shown as follows:

\[
T = t \ast F
\]

(19)
\[ y = \sin(2 \pi T f) \]  

where \( T \) is the time series (s), \( t \) is the sampling time (s), \( F \) is the sampling frequency (Hz), \( y \) is the stimulated data (cm s\(^{-1}\)), and \( f \) is the constant (0.05, -).
Figure 2. Schematic diagram of clean data and contaminated data of \( u \) direction with different degrees of outliers: (a) Clean data and contaminated data with 30\% outliers; (b) Clean data and contaminated data with 20\% outliers; (c) Clean data and contaminated data with 15\% outliers; (d) Clean data and contaminated data with 10\% outliers; (e) Clean data and contaminated data with 5\% outliers; (f) Clean data and contaminated data with 2\% outliers; (g) Clean data and contaminated data with 1\% outliers.
2.2.2. Field Measured Data

The field velocity data were measured in the penaeid shrimp (*Penaeus japonicus*) culture pond in Karatsu City of Saga Prefecture in Kyusyu District, Japan. The area of the pond was one hectare (length × width, 135 × 70 m). Figure 3 shows the paddle wheel aerators and the monitoring points of water current velocity. The maximum depth was 2.3 m. Eight aerators were individually set around the pond, and 44 measuring points were evenly distributed in the pond, with the interval of 26 m in transverse direction, and 14 m in longitudinal direction. The measuring depth was 0.2 m below the water surface, and extended downward every 0.3 m. An acoustic doppler velocimeter (VECTOR, NORTEK Inc., Akershus, Norway) was used with a sampling frequency of 16 Hz.

![Figure 3. The distribution of paddle wheel aerators and the monitoring points of water current velocity.](image)

2.3. Statistical Method

In order to test the effects of the three methods on detecting contaminated data with different degrees of outliers, this paper counts the correct detection rate (cdr) and over processing rate (opr) (Equations (21) and (22)) of each of the processed polluted data, and compares the kurtosis coefficient (Equation (23)) and skewness coefficient (Equation (24)) of the clean data and the contaminated data processed by all methods. The correct detection rate is the ratio of real outliers detected correctly by the algorithm. The over processing rate is the ratio of real values detected as outliers by the algorithm. Kurtosis coefficient is a characteristic number that is used to characterize the peak height of probability density distribution curve at the average value. High kurtosis indicates extreme differences in data that are greater or less than the average. Skewness coefficient characterizes the degree of asymmetry of the probability distribution density curve relative to the mean value. Detailed formulas can be described as:

\[
\text{cdr} = \frac{cdn}{on} \times 100\%,
\]  

(21)
where \( cdr \) is the correct detection rate (%), \( opr \) is the over processing rate (%), \( dn \) is the number of detected outliers \((-\)) , \( sn \) is the number of outliers \((-\)) , and \( cdn \) is the number of correctly detected outliers \((-\)) .

3. Results

3.1. Algorithm Verification

The total time of simulated sample data is 12,000 s. The number of detected outliers, correct detection rate and over processing rate by three algorithms are described in Table 1.

| Method          | Time (s) | Case | \( dn (-) \) | \( opr (%) \) | \( cdr (%) \) |
|-----------------|----------|------|---------------|---------------|---------------|
| Phase space     | 12,000   | 1    | 505           | 326.67        |               |
|                 |          | 2    | 906           | 287.08        |               |
|                 |          | 3    | 1828          | 226.50        |               |
|                 |          | 4    | 2765          | 169.08        |               |
|                 |          | 5    | 3457          | 134.50        |               |
|                 |          | 6    | 4108          | 114.67        |               |
|                 |          | 7    | 5106          | 85.97         |               |
| Robust estimation| 12,000   | 1    | 39            | 0.00          |               |
|                 |          | 2    | 82            | 0.00          |               |
|                 |          | 3    | 222           | 0.00          |               |
|                 |          | 4    | 417           | 0.00          | Figure 4      |
|                 |          | 5    | 631           | 0.00          |               |
|                 |          | 6    | 841           | 0.00          |               |
|                 |          | 7    | 1247          | 0.00          |               |
| 3DRPS           | 12,000   | 1    | 137           | 14.17         |               |
|                 |          | 2    | 260           | 8.75          |               |
|                 |          | 3    | 664           | 10.67         |               |
|                 |          | 4    | 1305          | 9.08          |               |
|                 |          | 5    | 1945          | 8.33          |               |
|                 |          | 6    | 2596          | 8.42          |               |
|                 |          | 7    | 3862          | 7.67          |               |

The values of \( dn \) and \( opr \) for PST in all cases are higher than those for RE and 3DRPS. RE and 3DRPS have less overtreatment phenomena, with the \( opr \) values as 0% and 9.58 ± 2.23%, respectively. The \( cdr \) of RE is the lowest. The values of \( cdr \) of all degrees are between 30–40%, the highest and lowest values of \( cdr \) are 37.00% (case 3) and 32.50% (case 1), and the average \( cdr \) is 34.73 ± 1.3%. The \( cdr \) of PST, with the average as 70.57 ± 16.72%, is greatly affected by the degree of pollution, in which two cases (cases 1 and 2) are higher than 90%, two cases (cases 3 and 4) are between 60–90%, and the remaining three cases are lower than 60%. For 3DRPS, \( cdr \) is higher than 99%, in which two cases are 100% (cases 1 and 3) with an average of 99.76 ± 0.17%. The \( cdr \) is significantly higher than the other two algorithms (\( F = 93.72, p < 0.001 \)) with the most stable detecting ability. Figure 4 shows the correct detection rates of three methods.
remaining three cases are lower than 60%. For 3DRPS, \( \text{cdr} \) is higher than 99%, in which two cases are significantly higher than the other. Figure 4 shows the correct detection rates of three methods.

Figures 5 and 6 are comparisons of kurtosis and skewness of processed data among three methods in three directions. The average kurtosis values of the clean data and the contaminated data are 1.507 and 6.182, respectively. The kurtosis values of PST, PE and 3DRPS are 1.911 ± 0.580, 1.882 ± 0.322, and 1.507 ± 0.000, respectively. Compared with kurtosis of contaminated data, the decreases of the three methods are 69.09%, 69.55% and 75.62%, respectively. According to Figure 5, 21 groups (including three dimensions) could be found closest to the kurtosis coefficient of clean data in 3DRPS, followed by PST with 12 groups and RE with 8 groups. The average skewness of clean data and contaminated data are –0.068 and –0.038, respectively. The skewness coefficients of PST and RE are –0.066 ± 0.008 and –0.067 ± 0.007. Skewness of 3DRPS is close to skewness of clean data, which is –0.068 ± 0.000. Compared with skewness of contaminated data, the decreases of three methods are 75.12%, 79.24% and 80.03%, respectively. The 3DRPS with 21 groups shows the closest skewness to the clean data, then followed by PST with 14 groups and RE with 13 groups.

![Figure 4. Comparisons of correct detection rates among three methods.](image1)

![Figure 5. Cont.](image2)
Figure 5. Comparisons of kurtosis of processed data among three methods: (a) kurtosis of processed data of $u$ direction; (b) kurtosis of processed data of $v$ direction; (c) kurtosis of processed data of $w$ direction.

Figure 6. Cont.
Figure 6. Comparisons of skewness of processed data among three methods: (a) skewness of processed data of \( u \) direction; (b) skewness of processed data of \( v \) direction; (c) skewness of processed data of \( w \) direction.

### 3.2. Algorithm Application

After the algorithm verification, three methods were applied to the monitoring data of a shrimp pond. The comparison data of the detection results are shown in Figure 7 and the analysis results are shown in Table 2. The statistical parameter of field measured data show that the kurtosis of raw data is 26.06, 25.53 and 25.68, respectively. Kurtosis of \( x \) and \( y \) dimension data processed by phase space decreased slightly (about 23% and 12%) and that of \( z \) dimension decreased by 45%. Kurtosis of the data processed by RE and 3DRPS decreased by about 87% and 90%, respectively.

![Figure 7. Cont.](image-url)
With the increase of the degree of pollution, the choice of COR in 3DRPS is 1.483. The reason for taking 2.5 for cutoff values is that the probability is not far from the specified range and can be regarded as a special approximation. Therefore, the choice of COR comes from a normal distribution. When data decreases significantly. When data comes from a normal distribution, the correction coefficient used in the RE method is 1.483 [19]. Parsheh et al., (2010) pointed out that based on the standard deviation of the probability density function, the optimal selection range for the COR is 1.25–1.45, and 1.35 is highly suggested [23]. The 1.483 proposed by Rousseuw [19] is not far from the specified range and can be regarded as a special approximation. Therefore, the choice of COR in 3DRPS is 1.483. The reason for taking 2.5 for cutoff values is that the probability of $|z_i| > 2.5$ is small when there are no outliers and $x_i$ comes from a normal distribution [19]. In the 3DRPS method, 2.5 is taken as the cutoff value to maintain consistency with the RE method. The influence of the selection of COR and cutoff value on the detection effect can be further studied.

According to the algorithm verification results, it is clearly stated that PST is effective for data with less contamination. With the increase of the degree of pollution, $cdr$ decreases significantly. When data

**Table 2. Statistical parameter values of monitoring velocity data.**

| Methods       | Dimensions | Average  | Standard Deviation | Kurtosis  | Skewness  | Max     | Min     |
|---------------|------------|----------|--------------------|-----------|-----------|---------|---------|
| Raw Data      | x          | -0.02067559 | 0.274184           | 26.0623  | -0.4404   | 2.229   | -2.467  |
|               | y          | -0.009378801 | 0.256644           | 25.52965 | 0.165199  | 2.408   | -2.16   |
|               | z          | -0.001662331 | 0.049877           | 25.68186 | 0.242418  | 0.464   | -0.491  |
| Phase Space   | x          | -0.01751828  | 0.142615           | 20.16426 | -0.03063  | 1.174   | -1.204  |
|               | y          | -0.01292775  | 0.126272           | 22.3524  | 0.154246  | 1.097   | -1.12   |
|               | z          | -0.001159016 | 0.029374           | 14.01004 | 0.628032  | 0.213   | -0.19   |
| Robust Estimation | x     | -0.0201469   | 0.086652           | 2.67489  | -0.020317 | 0.2     | -0.234  |
|               | y          | -0.0139824   | 0.075512           | 3.211917 | -0.02248  | 0.181   | -0.203  |
|               | z          | -0.002055034 | 0.015104           | 2.976916 | -0.04973  | 0.035   | -0.039  |
| 3DRPS         | x          | -0.01963378  | 0.085787           | 2.697075 | -0.21987  | 0.2     | -0.234  |
|               | y          | -0.01401899  | 0.074861           | 3.26277  | -0.02469  | 0.181   | -0.203  |
|               | z          | -0.00218289  | 0.014823           | 3.043208 | -0.03984  | 0.035   | -0.039  |

4. Discussion

The correction coefficient used in the RE method is 1.483 [19]. Parsheh et al., (2010) pointed out that based on the standard deviation of the probability density function, the optimal selection range for the COR is 1.25–1.45, and 1.35 is highly suggested [23]. The 1.483 proposed by Rousseuw [19] is not far from the specified range and can be regarded as a special approximation. Therefore, the choice of COR in 3DRPS is 1.483. The reason for taking 2.5 for cutoff values is that the probability of $|z_i| > 2.5$ is small when there are no outliers and $x_i$ comes from a normal distribution [19]. In the 3DRPS method, 2.5 is taken as the cutoff value to maintain consistency with the RE method. The influence of the selection of COR and cutoff value on the detection effect can be further studied.

According to the algorithm verification results, it is clearly stated that PST is effective for data with less contamination. With the increase of the degree of pollution, $cdr$ decreases significantly. When data...
contain more than 10% outliers, the value of $cdr$ decreases slowly, which is basically consistent with the previous studies. Islam et al. (2013) proposed that the PST performs well when the sample data contain 5% outliers [17]. Jesson et al. (2013) pointed out PST is more suitable for data with fewer outliers in laboratory or field experiments (the calculation of detection threshold of phase space is based on the mean value and standard deviation, which are sensitive to outliers) [24]. The existence of extreme outliers can significantly affect the performance of the two parameters; the positioning value and the scale estimation value will subsequently move far away from real data and increase the detection threshold, causing the outliers to be regarded as real data and miss elimination [21]. Although this problem can be improved by iteration, the first and second derivatives of velocity data near outliers may exceed the ellipsoid boundary, hence some real data around outliers will be determined as outliers [23,25]. In addition, PST has no effect on continuous outlier data [14]. The $cdr$ of RE has little correlation with the degree of pollution, while with low Gaussian distribution efficiency, its detection efficiency is negatively correlated with sample size [26]. Robust estimation calculates the detection threshold by the median of all absolute deviations; therefore, it can avoid iteration and overtreatment of data process. The breakdown point of ADV is 50%, indicating RE is not suitable when over half of the data are contaminated. In contrast, the 3DRPS method has the highest $cdr$ due to preprocessing sample data. Based on the principle of robust estimation, the dynamic window method separates the sample data into windows, then calculates the robust positioning values and scale estimation values in the window according to the located data. When $O_i > 0$ (Equation (4)), the correlated data are deleted, which significantly improves the detection accuracy of real data. After the first filtration, most outliers and continuous outliers are eliminated and their impacts on average and standard deviation will be reduced. Then PST is used for secondary filtration, the iteration due to improving detection accuracy can be avoided. And thus it greatly reduces the overtreatment phenomenon and avoids the impact on the original data structure. In previous studies, data for algorithm validation usually contain no more than 25% outliers [21,23]. In this paper, the highest data pollution degree is 30%, and the result shows 3DRPS can be effectively applied to data with 30% outliers.

Coefficient of kurtosis is a measure of the “peakedness” of the probability distribution of a variable. In general, the normal distribution is 3 [27]. The coefficient of skewness is a measure of symmetry of the probability distribution, so when the distribution is symmetrical, the skewness coefficient is zero. If the skewness coefficient is greater than 0, the distribution will be right-biased, and the distribution is left-biased when the skewness coefficient is less than 0 [28]. It can be seen from Figures 5 and 6 that the simulated clean data present a cosine curve, which is not normally distributed and not completely symmetrical. The kurtosis coefficient of the cleaning data is 1.5 and the skewness coefficient is −0.067. Therefore, it is necessary to compare the kurtosis and skewness of the data detected by the algorithm with the kurtosis and skewness of the clean data to reflect the restoration of the algorithm on the data. In Figure 6, the skewness of $u$ and $w$ in contaminated data are negative values, while that of $v$ is a positive value, which shows no regularity. The skewness is affected by the position and the values of outliers. According to the kurtosis and the skewness of simulated data and processed data, the methods of PST and RE are less effective in processing extreme outliers and data symmetry. For kurtosis, although the number of degrees of PST are closer to clean data than that of RE, the kurtosis of data processed by phase space are significantly higher when outliers increase. The main reason is that the extreme outliers cannot be detected completely because of the weakness of PST for the increase of outliers. Though the $cdr$ of RE is lower than PST, it runs better in detecting extreme outliers, causing lower kurtosis of data by RE. Skewness of PST and RE are not influenced by the degree of outliers. Furthermore, RE only calculates symmetric statistics about the position estimator, and can hardly deal with the skewness [29]. The 3DRPS performs well in dealing with the extreme outliers and the symmetry of contaminated data, and its parameter values show no difference from those of clean data. It is obviously shown that outliers of field velocity monitoring data processed by 3DRPS can be removed and the kurtosis and skewness can be greatly reduced; moreover, the structure of the field data are presented closer to a real one.
The PST and RE are both described for the application of the time series of a single component of velocity. Most ADV probes report three-dimensional velocities. The reported orthogonal velocity components are interdependent, since the instrument actually measures beam velocities along the bistatic axes of the sending and receiving acoustic elements, and it converts these velocities to orthogonal velocity components via multiplication by a transformation matrix [30]. The factors that produce outliers often affect only one of the individual beam velocities; while after multiplication by the transformation matrix, the single affected beam velocity taints all three of the orthogonal velocity components [29]. In the process of algorithm verification and application, 3DRPS is successful in three-dimensional detection, and the associated data in all three-time series will be deleted when an outlier is identified in any one of time series, even if the outliers do not appear in other time data series.

It can be seen from Figure 7 that in the measured data of the shrimp pond, outliers mainly appear between 2000 s and 7000 s and there are some outliers near 10,000 s. After 3DRPS detection, these outliers were obviously detected, and the data after detection were more centralized to the mean value. The maximum values of data in three directions before and after detection are 2.229, 2.408, 0.464 and 0.02, 0.181, 0.035, respectively. The minimum values of data are −2.467, −2.16, −0.491 and −0.234, −0.203, −0.039, respectively.

5. Conclusions

In order to improve the performance of monitoring velocity data that contain large amounts of outliers from the paddle wheel aerator in the shrimp pond, a new method named Three Dimensional Rousseeuw Phase-Space Method through two-step filtration and three-dimensional simultaneous detection was developed, and it was compared with two widely-used outlier detecting methods (Robust Estimation Method and Phase Space Thresholding Method). The results show that 3DRPS approach (1) has a higher detection accuracy and lower over-processing of data, (2) is suitable for highly contaminated data (30%), and (3) is close to the original data after detection. This implied 3DRPS is an effective method for despiking water flow data and could provide accurate velocity data for the studies of the velocity distribution and the estimation of bottom friction. However, the amount of sample data in this study is relatively inadequate, indicating the performance of 3DRPS for a large amount data needs to be validated in future studies.

Author Contributions: Conceptualization, J.Z.; methodology, J.Z.; data analysis, C.Z., F.Y. and S.Z.; experiment, J.Z. and D.K.; resources, F.Y. and S.Z.; writing—Original draft preparation, C.Z. and J.Z.; writing—Review and editing, J.Z., C.Z. and D.K.; supervision, R.W. and D.K. All authors have read and agreed to the published version of the manuscript.

Funding: This study is partly supported by the Young Orient Scholars Program of Shanghai (No. QD2017038), and the National Natural Science Foundation of China (No. 41807341).

Acknowledgments: We are deeply indebted to a shrimp aquaculture company for providing a pond for field investigation. We also thank Yamayoshi and Chiba for supporting the current velocity measurement.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fishery and Aquaculture Statistic 2016; FAO Yearbook; Food and Agriculture Organization of the United Nations: Rome, Italy, 2018; p. 104.
2. The State of World Fisheries and Aquaculture 2018–Meeting the Sustainable Development Goals; Food and Agriculture Organization of the United Nations: Rome, Italy, 2018; p. 20.
3. Megraw, W.; Teichert-Coddington, D.R.; Rouse, D.B.; Boyd, C.E. Higher minimum dissolved oxygen concentrations increase penaeid shrimp yields in earthen ponds. Aquaculture 2001, 199, 311–321. [CrossRef]
4. Duan, H.B.; Zhang, H.; Huang, Q.F.; Zhang, Y.; Hu, M.; Niu, Y.; Zhu, J. Characterization and environmental impact analysis of sea land reclamation activities in China. Ocean Coast. Manag. 2016, 130, 128–137. [CrossRef]
5. Kitazawa, D.; Zhang, J. Measurement of water current field created by paddle wheel aerators in shrimp culture pond. In Proceedings of the IEEE Oceans, Shanghai, China, 10–13 April 2016; pp. 1–4.
6. Boyd, C.E.; Torrans, E.L.; Tucker, C.S. Dissolved Oxygen and Aeration in Ictalurid Catfish Aquaculture. J. World Aquacult. Soc. 2017, 49, 7–70. [CrossRef]
7. Voulgaris, G.; Trowbridge, J.H. Evaluation of the Acoustic Doppler Velocimeter (ADV) for Turbulence Measurements. J. Atmos Ocean. Tech. 1998, 15, 272–289. [CrossRef]
8. Sharma, A.; Maddirala, A.K.; Kumar, B. Modified singular spectrum analysis for despiking acoustic Doppler velocimeter (ADV) data. Measurement 2018, 117, 339–346. [CrossRef]
9. Hejazi, K.; Falconer, R.A.; Seifi, E. Denoising and Despiking ADV Velocity and Salinity Concentration Data in Turbulent Stratified Flows. Flow Meas. Instrum. 2016, 52, 83–91. [CrossRef]
10. McLelland, S.J.; Nicholas, A.P. A new method for evaluating errors in high–frequency ADV measurements. Hydrol. Process. 2000, 14, 351–366. [CrossRef]
11. Fugate, D.C.; Friedrichs, C.T. Determining concentration and fall velocity of estuarine particle populations using ADV, OBS and LISST. Cont. Shelf Res. 2002, 22, 1867–1886. [CrossRef]
12. Li, H.; Wang, P.; Yang, C. Measurement mechanism and error control of acoustic Doppler flow meter. Natl. Symp. Hydrodyn. 2007, 407–410.
13. Mori, N.; Suzuki, T.; Kakuno, S. Noise of Acoustic Doppler Velocimeter Data in Bubbly Flows. J. Eng. Mech. 2007, 133, 122–125. [CrossRef]
14. Goring, D.G.; Nikora, V.I. Despiking Acoustic Doppler Velocimeter Data. J. Hydraul. Eng. 2002, 128, 117–126. [CrossRef]
15. Cea, L.; Puertas, J.; Pena, L. Velocity measurements on highly turbulent free surface flow using ADV. Exp. Fluids 2007, 42, 333–348. [CrossRef]
16. Mohammad, V.; Kumars, M.; Maryam, A. Detection of Outlier in 3D Flow Velocity Collection in an Open–Channel Bend Using Various Data Mining Techniques. Iran. J. Sci. Technol. Trans. Civ. Eng. 2019, 43, 197–214.
17. Islam, M.R.; Zhu, D.Z. Kernel density–based algorithm for despiking ADV data. J. Hydraul. Eng. 2013, 139, 785–793. [CrossRef]
18. Rousseeuw, P.J.; Leroy, A.M. A robust scale estimator based on the shortest half. Stat. Neerl. 1988, 42, 103–116. [CrossRef]
19. Rousseeuw, P.J. Robust Estimation and Identifying Outliers. In Handbook of Statistical Methods for Engineers and Scientists, 2nd ed.; McGraw-Hill: New York, NY, USA, 1998; pp. 16.2–16.4.
20. Nikora, V.I.; Goring, D.G. ADV Measurements of Turbulence: Can We Improve Their Interpretation? J. Hydraul. Eng. 1998, 124, 630–634. [CrossRef]
21. Wahl, T.L. Discussion of “Despiking Acoustic Doppler Velocimeter Data” by Derek, G. Goring and Vladimir, I.Nikora. J. Hydraul. Eng. 2003, 129, 484–487. [CrossRef]
22. Zou, H.; Xiang, D.W.; Jing, Y.G. Data smoothing methods for DVL. Tech. Acoust. 2008, 27, 507–510.
23. Parsheh, M.; Sotiropoulos, F.; Portéagel, F. Estimation of Power Spectra of Acoustic–Doppler Velocimetry Data Contaminated with Intermittent Spikes. J. Hydraul. Eng. 2010, 136, 368–378. [CrossRef]
24. Jesson, M.; Sterling, M.; Bridgeman, J. Despiking velocity time–series—Optimisation through the combination of spike detection and replacement methods. Flow Meas. Instrum. 2013, 30, 45–51. [CrossRef]
25. Patel, A.; Khatua, K.K. Reliability Study in Despiking of ADV Data to Open Channel Flow; L. D. College of Engineering: Ahmedabad, India, 2017.
26. Rousseeuw, P.J.; Croux, C. Alternatives to median absolute deviation. J. Am. Stat. Assoc. 1993, 88, 1273–1283. [CrossRef]
27. Andrea, B. Robust measures of skewness and kurtosis for macroeconomic and financial time series. Appl. Econ. 2020, 52, 637–670.
28. Ömer, K. Various filtering algorithms used to eliminate outliers in velocity time series obtained by ADVs (acoustic Doppler velocimeter). Arab. J. Geosci. 2013, 6, 2691–2697.
29. Razaz, M.; Kawanisi, K. Signal post-processing for acoustic velocimeters: Detecting and replacing spikes. *Meas. Sci. Technol.* **2011**, *22*, 125404. [CrossRef]

30. Kraus, N.C.; Lohrmann, A.; Cabrera, R. New Acoustic Meter for Measuring 3D Laboratory Flows. *J. Hydraul. Eng.* **1994**, *120*, 406–412. [CrossRef]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).