Outlier detection and sequence reconstruction in continuous time series of ocean observation data based on difference analysis and the Dixon criterion

Jiang Jingang,1,2* Sun Lu,1 Fan Zhongya,3,4 Qi Jiaguo1
1Ocean College, Zhejiang University, Hangzhou, Zhejiang, China
2Institute of Remote Sensing and Earth Sciences, Hangzhou Normal University, Hangzhou, Zhejiang, China
3State Key Laboratory of Estuarine and Coastal Research, ECNU, Shanghai, China
4South China Institute of Environmental Sciences MEP, Guangzhou, Guangdong, China

Abstract

In light of the specific characteristics of the long-term record and complex sequence nature of the ocean observation data, a new method was developed based on the original Dixon detection criteria to be specifically detect and remove data outliers. This method combines the two traditional methods of data quality control and Dixon detection theory and assumes that the second-order differential sequence of parameter measurements passes an appropriate stationarity test. Thus, the measurement attributes are considered to be in the same physical state and to occupy a small range of time and space, equivalent to a parallel observation test. Provided that the observations over a small range of time and space correspond to the record of a sequence covering a short period of time, this short time sequence is treated as a sliding window in the proposed new method. Outliers are detected based on lookup-table after an index parameter \(Q\) is calculated within the sliding window. A correlation analysis and the test results show that the proposed new method can effectively instantiate a sequence of outliers characterized by different phases. Compared with other existing methods, the new method proved to be computationally efficient and easily programmable for practical implementation. Further, this method preserves the original data because the outliers are replaced by an inverse distance-weighted average of the recorded data within the window, while other data were intact.

Because resources in terrestrial ecosystems are limited, the efficient use of ocean resources is extremely important for economic development and carbon-reduction policies (Pan et al. 2012). In recent years, the speed with which marine resources have been exploited is much greater than ever before. The capacity for ocean and coastal monitoring has been significantly improved to more efficiently manage human activities in marine resource development and natural marine disasters. However, the quality of ocean observations must be ensured to provide accurate information for operational management practices. The control and improvement of data quality are the backbone of meaningful analysis and information extraction in business and scientific research (Shi et al. 2000). Similarly, quality control is an important part of ocean data assimilation systems. If erroneous data are assimilated, they can cause immediate spurious disruptive overturning and error propagation because of the sparse distribution of oceanographic data (Ingleby and Huddleston 2007; Cosme et al. 2010).

Presently, global ocean science has entered a multidisciplinary and three-dimensional age. An integral marine integrated monitoring system consists of three parts: marine environment integrated observation, digital communication and management, and data processing and applications. The forms of ocean observation can be remote sensing of satellites and aircrafts and in situ observation of scientific research ships, offshore fixed stations and buoy stations (Fig. 1). Offshore fixed-station observation is a key approach to marine monitoring in China. This approach can continuously monitor principal parameters such as the sea waves, sea surface temperature (SST), tide, storm surge, and marine meteorological conditions. In the 3 million square kilometers of sea area of China, survey projects are primarily implemented via ocean-observing buoys and ocean research projects. Since the end of 2006, the State Oceanic Administration and some government departments have paid increased attention to the construction and improvement of offshore fixed stations, island stations, offshore platforms, ships

*Correspondence: gangzg@163.com

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
and buoy stations. According to incomplete statistics from early 2009, nearly 100 fixed offshore automatic observatories and nearly 100 marine automatic observation systems have been brought on line (Zhu 2009). Three-dimensional and continuous time series observation has become an active area of developmental marine observation.

Therefore, the implementation of marine data quality control has become a notably important problem in the assurance of sea information reliability. The common outlier detection methods are the inspection of extreme values, judgment of consistency, increasing scientific rationalization of messy data, Grubbs’ test and Dixon inspection (Chen 1991; Huang et al. 1999; Shi 2010). In the early 1960s, the United States reorganized existing marine data and established a marine data processing and quality control system. During the Eighth Five-Year Plan, the CMA (China Meteorological Administration) built a Meteorological Data Processing and Application Service System, in which global ocean observation data from 1854 to 1995 were collected. These data were processed in a unified manner through the preliminary quality control method, and global ocean climate data sets were established (Ma et al. 1999). Data quality control is based on both objective criteria and human experience (Doong et al. 2007). James (1993) presented a method to detect outliers in a time series. This method was based on the principle that the deviations of measurements from their mean value should smoothly vary and follow a uniform distribution. The proposed cutoff was the outlier fence, which is defined by three times the standard deviation of the time series measurements. These fences established the upper and lower limits that could be used to check for data outliers (McKinney 1993). Ingleby and Huddleston (2007) used the 1956–2004 archive data and Bayesian probability theory to achieve quality control of ocean temperature and salinity profiles (Ingleby and Huddleston 2007).

The difficulties that hamper the elimination of uncertainty in measurements are determined by the stochastic properties of the observations. The ideal research technique to confront this type of random phenomenon is data analysis, including statistics, the use of which can reduce the uncertainty of measurement data (Dou et al. 2012). Based on the regional and seasonal stability of the water masses, Huang et al. (1993) used probabilistic and statistical methods to control the hydrological data quality according to the normality of the probability distributions of marine hydrological factors. Since its initiation in 2000, the international Array for Real-time Geostrophic Oceanography (ARGO) program has paid strict attention to data quality control. Two data quality control methods have been proposed by the ARGO program: one method based on the statistical characteristics of historical data and another based on a model that relates temperature to salinity (Wong et al. 2003; Ji et al. 2004; Wang et al. 2012).

With the progress and development of marine environmental observation technology, in situ and continuous time series observation will represent the cutting-edge methodological wavefront for future data surfers. Increasing amounts of marine environmental observation data are challenging the existing data-processing and quality control methods. To
solve this problem, this paper proposes a new method based on the original Dixon detection criteria for outlier detection and sequential reconstruction with particular reference to Dou et al. (2012). The method is developed according to the notion that the changes in the sequence of differences calculated from the measured parameters is consistent with the hypothesis of stationarity. The analysis is conducted using stationarity tests on random sequences and various statistical properties of time series data. The analysis results indicate that the proposed method can effectively identify outlier information in continuous time series, and the observation method is effective for improving the quality of marine observation data.

**Materials and methods**

**Data sources**

The ocean is an important component of the earth-sea-air system. It is also a complex physical, chemical, and ecological system. Thus, the subject matter and disciplinary span of ocean research and observation are notably extensive; they include the observation of marine meteorological parameters, observation of various physical processes in the ocean at various scales, and data acquisition on water quality and environmental elements. Table 1 shows the experimental data sources, items, measuring instruments and their accuracies, sampling period and sampling locations in this paper. The selected data for this paper include observations of ocean physical processes, meteorological observations, and data on environmental parameters related to water quality.

As shown in Table 1, the measured parameters are the water level and velocity, sea surface atmospheric pressure, dissolved oxygen (DO), and sea surface temperature. The shortest sampling period is 1 min, and the longest is 1 d.

**Differential characteristics of continuous time series observation data**

Marine monitoring can be divided into two sectors: underwater monitoring and above-water monitoring. The commonly used instruments for underwater monitoring are acoustic Doppler current profilers (ADCPs), conductivity–temperature–depth (CTD) sensors, the GO8050 barometer system, and multi-parameter water-quality-monitoring devices. The principal parameters monitored are associated with or given by the currents, turbulence, turbidity, water quality and carbon dioxide pressure (PCO2). The commonly used instruments for above-water monitoring are buoys, high-frequency surface wave radars (HFSWRs), and ocean satellites. The principal parameters monitored are associated with the waves, weather, and hydrology (Li 2012). To detect inharmonious information in continuous time series observation data, this paper selects two outlier detection methods: differential analysis and the Dixon detection criterion. The underwater test data in this paper include water level and velocity data, which were obtained with the Aanderaa RCM-9 current meter; atmospheric pressure data, which were obtained with the General Oceanics GO8050 barometer system; and dissolved oxygen data, which were obtained with the Hydrolab multi-parameter DSSX water quality monitor. The above-water test data used in this paper are the SST data obtained with the MODIS satellite. The Ocean Biology Processing Group (OBPG) provides Aqua MODIS SST level 3 daily product data (https://oceancolor.gsfc.nasa.gov/). In this data set, the highest sampling period is for the dissolved oxygen measurements with an interval of 1 min. The SST measurements have the lowest sampling period in the data set, with an interval of 1 d.

The analysis in this study used the Augmented Dickey-Fuller (ADF) statistic (Hamilton 1994; Greene 2003) method to test whether a time series is stationary.

The ADF test was developed based on the Dickey-Fuller (DF) test (Dickey and Fuller 1979). The DF test is a statistical test to check for stationarity. The null hypothesis indicates that the time series is non-stationary. The test results comprise a Test Statistic and some Critical Values for difference confidence levels. If the “Test Statistic” is less than the “Critical Value,” the null hypothesis can be rejected, indicating that the series is stationary. The Dickey-Fuller test can be used to test an autoregressive model for the presence of unit roots.

**Table 1.** Information about the data sources.

| Item                | Instrument                        | Instrument accuracy | Sampling period | Sampling location          |
|---------------------|-----------------------------------|---------------------|-----------------|----------------------------|
| Water level (m)     | Aanderaa Current Meter (RCM-9)    | <0.01 m             | 5 min           | 22°41′32.55″N               |
|                     |                                   |                     |                 | 113°9′59.80″E               |
| Velocity (cm/s)     | Aanderaa Current Meter (RCM-9)    | ±0.15 cm/s          | 5 min           | 22°41′32.55″N               |
|                     |                                   |                     |                 | 113°9′59.80″E               |
| Atmosphere (hPa)    | GO8050 Barometer                  | ±1 hPa              | 2 min           | 26°55′6.96″N                |
|                     |                                   |                     |                 | 120°12′28.80″E              |
| DO (mg/L)           | Hydrolab Multi-parameter Water Quality Monitor (DSSX) | ±0.1 mg/L (<8 mg/L); ±0.2 mg/L (>8 mg/L); | 1 min | 30°39′6.24″N |
|                     |                                   |                     |                 | 121°20′43.08″E              |
| SST (°C)            | MODIS SST Level 3 products        | <0.5 °C             | 1 d             | 30°25′35.58″N               |
|                     |                                   |                     |                 | 123°34′33.09″E              |

Note: MODIS sea surface temperature data (SST, °C) were processed using the DINEOF method (Beckers and Rixen 2003; Alvera-Azcárate et al. 2005) to interpolate the missing pixels in the remote sensing data.
Consider a simple general AR(1) process given by:

\[ Y_1 = \phi_1 Y_{t-1} + \epsilon_t \]  

(1)

The characteristic equation for the time series \( Y \) is:

\[ \lambda - \phi_1 = 0 \]  

(2)

When the characteristic root \( \phi_1 \) is in the unit circle, the sequence is stationary; otherwise, the sequence is non-stationary. Thus, we can test the stationarity of the sequence by checking whether the characteristic root is inside or outside the unit circle, and this test is called the unit root test.

Simultaneously subtracting \( Y_{t-1} \) in Eq. 1 and a from both sides gives:

\[ Y_1 - Y_{t-1} = (\phi_1 - 1) Y_{t-1} + \epsilon_t \]  

(3)

The DF test is equivalent to the following test:

\[ H_0 : \rho = 0 \iff H_1 : \rho < 0 \]

\[ \rho = \phi_1 - 1 \]  

(4)

The DF test statistic model is:

\[ \tau = \frac{\hat{\rho}}{S(\hat{\rho})} \]  

(5)

where \( S(\hat{\rho}) \) is the sample standard deviation of parameter \( \rho \); \( \hat{\rho} \) is the method of ordinary least squares (OLS) for parameter \( \rho \). The DF test is a one-sided test when the significance level is assumed to be \( \alpha \). When \( \tau \leq \tau_\alpha \), we reject the null hypothesis that the sequence is significantly stationary; otherwise, we accept the null hypothesis, i.e., the sequence is non-stationary.

The DF test is only suitable for the one-order stationarity test autoregressive process. To make the DF test applicable to the AR(\( p \)) stationary test process, ADF is needed to make some corrections to the DF test.

The AR(\( p \)) process is given by:

\[ Y_1 = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t \]  

(6)

The characteristic equation for the time series \( Y \) is:

\[ \lambda^p - \phi_1 \lambda^{p-1} - \cdots - \phi_p = 0 \]

\[ \iff \lambda - \phi_1 \lambda - \phi_2 \cdots - \phi_p = 0 \]

\[ \Rightarrow \phi_1 + \phi_2 + \cdots + \phi_p = 1 \]  

(7)

We make the following hypothesis test:

\[ H_0 : \rho = 0 \iff H_1 : \rho < 0 \]

\[ \rho = \phi_1 + \phi_2 + \cdots + \phi_p - 1 \]  

(8)

The three types of model variant are as follows:

**AR:** Autoregressive model variant, which specifies a test of the null model.

\[ Y_1 = Y_{t-1} + \beta_1 \Delta Y_{t-1} + \beta_2 \Delta Y_{t-2} + \cdots + \beta_p \Delta Y_{t-p} + \epsilon_t \]  

(9)

against the alternative model

\[ Y_1 = \phi Y_{t-1} + \beta_1 \Delta Y_{t-1} + \beta_2 \Delta Y_{t-2} + \cdots + \beta_p \Delta Y_{t-p} + \epsilon_t \]  

(10)

where \( \Delta \) is the differencing operator such that \( \Delta Y_1 = Y_1 - Y_{t-1} \).

The number of lagged difference terms \( p \) is specified by the user. \( \epsilon_t \) is a mean zero innovation process.

With, AR(\( I \)) coefficient \( \phi < 1 \).

**ARD:** Autoregressive model with drift variant, which specifies a test of the null model.

\[ Y_1 = \phi Y_{t-1} + \beta_1 \Delta Y_{t-1} + \beta_2 \Delta Y_{t-2} + \cdots + \beta_p \Delta Y_{t-p} + \epsilon_t \]  

(11)

against the alternative model

\[ Y_1 = c + \phi Y_{t-1} + \beta_1 \Delta Y_{t-1} + \beta_2 \Delta Y_{t-2} + \cdots + \beta_p \Delta Y_{t-p} + \epsilon_t \]  

(12)

with drift coefficient \( c \) and AR(\( I \)) coefficient \( \phi < 1 \).

**TS:** Trend-stationary model variant, which specifies a test of the null model.

\[ Y_1 = c + y_{t-1} + \beta_1 \Delta Y_{t-1} + \beta_2 \Delta Y_{t-2} + \cdots + \beta_p \Delta Y_{t-p} + \epsilon_t \]  

(13)

against the alternative model

\[ Y_1 = c + \delta t + \phi Y_{t-1} + \beta_1 \Delta Y_{t-1} + \beta_2 \Delta Y_{t-2} + \cdots + \beta_p \Delta Y_{t-p} + \epsilon_t \]  

(14)

with drift coefficient \( c \), deterministic trend coefficient \( \delta \), and AR(\( I \)) coefficient \( \phi < 1 \).

The test statistic value of the original observation data and second-order difference of the observation data \( (\alpha = 0.05) \) were used in the “adftest” function in MATLAB. As shown in Table 2, the high \( p \)-values \( (p > 0.05) \) of all original observation data indicate that it fails to reject the unit-root null. The results also show that the second-order difference of the observation data is stationary. Figure 2 shows the water level, velocity, and second-order differences of these quantities, as reported by an Aanderaa (RCM-9) current meter. The physical meaning of these results is that over a small range of time and space, the patterns of change in the observations are stable. In terms of practical quality control for continuous ocean observations, the use of various data acquisition methods, e.g., buoy observations and remote sensing observations, in comparisons involving differences and consistency is an important approach. However, this approach is costly, raises issues of spatial and temporal scales, and is difficult to implement. The stationarity of the second-order differences enables us to assume that over a small range of time and space (a small window), the observations are equivalent to those generated by parallel observation experiment.
Table 2. Comparison of the ADF test parameters between the original observation data and the second-order difference of the observation data ($z=0.05$).

| Item            | Original data |                   | Second-order difference data |                   |
|-----------------|---------------|-------------------|-------------------------------|-------------------|
|                 | p-values      | Test statistics   | Critical values               | p-values          | Test statistics   | Critical values   |
| Water level (m) | 0.19          | $-1.25$           | $-1.94$                       | 0.0010            | $-49.71$          | $-1.94$           |
| Velocity (cm/s) | 0.31          | $-0.94$           | $-1.94$                       | 0.0010            | $-42.65$          | $-1.94$           |
| Atm press (hPa) | 0.99          | 2.57              | $-1.94$                       | 0.0010            | $-85.43$          | $-1.94$           |
| DO (mg/L)       | 0.41          | $-0.65$           | $-1.94$                       | 0.0010            | $-31.16$          | $-1.94$           |
| SST (°C)        | 0.16          | $-1.36$           | $-1.94$                       | 0.0010            | $-117.21$         | $-1.94$           |

Dixon detection criterion and method to reconstruct outliers

The Dixon detection method, which was introduced by the mathematician Dixon in 1950, is one of the most popular tools for data quality control (Lei 1997; Sha 2003). The method does not need to estimate the mean and standard deviation. It identifies the gross error relative to the size difference according to the measured data. The method was also used to perform conformance checks and eliminate abnormal values of parallel test measurement data. The method’s calculations involve the following steps:

First, sort a set of $n$ measurements $x_1, x_2, x_3, \ldots, x_n$ into $x'_1, x'_2, x'_3, \ldots, x'_n$ from lowest to highest.

Second, consider the lowest and highest value in the ranked data sequence as the questionable values. Identify the lowest or highest value as outliers according to the following formula:

The $Q$ values below are the index parameters for making decisions about the outliers. We calculate:

\[
Q_{10} = \frac{x'_n - x'_{n-1}}{x'_n - x'_1}, \quad x'_n \text{ is a dubious value,} \quad 3 \leq n \leq 7
\]

\[
Q_{11} = \frac{x'_n - x'_{n-2}}{x'_n - x'_3}, \quad x'_n \text{ is a dubious value,} \quad 8 \leq n \leq 10
\]

\[
Q_{21} = \frac{x'_n - x'_{n-2}}{x'_n - x'_4}, \quad x'_n \text{ is a dubious value,} \quad 11 \leq n \leq 13
\]

\[
Q_{22} = \frac{x'_n - x'_{n-2}}{x'_n - x'_5}, \quad x'_n \text{ is a dubious value,} \quad 14 \leq n \leq 20
\]

After calculating the value of $Q$, threshold $Q_a$ is read from Table 3 according to the selected significance level $z$ and repeated testing time $n$. The decision to retain or eliminate the dubious data is made according to the following criteria:

1. If $Q > Q_{0.01}$, then the dubious values are outliers that must be processed.

Fig. 2. Features of the second-order differences of the continuous time series observation data.
Table 3. Decision table for Dixon criterion detection threshold.

| Significance level | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|-------------------|----|----|----|----|----|----|----|----|----|----|----|----|----|
| \( Q_{0.05} \)    | 0.642 | 0.562 | 0.507 | 0.554 | 0.512 | 0.477 | 0.575 | 0.546 | 0.521 | 0.546 | 0.524 | 0.505 | 0.489 |
| \( Q_{0.01} \)    | 0.782 | 0.698 | 0.637 | 0.681 | 0.635 | 0.597 | 0.674 | 0.642 | 0.617 | 0.640 | 0.618 | 0.597 | 0.580 |

2. If \( Q_{0.05} < Q < Q_{0.01} \), then the dubious values are deviant values that can be retained or processed. If it is necessary to address a dubious value, the median of the remaining valid data is commonly used in lieu of the abnormal data. However, an appropriate interpolation method such as spline interpolation may be more reasonable. In this paper, the inverse distance weighted average method is used instead of the median.

3. If \( Q < Q_{0.05} \), then the dubious values are normal values and should be retained.

4. Because it is difficult to perform parallel repeated observations in the marine continuous observation context, we must apply an appropriate transformation when using the Dixon detection criterion for outliers.

Dou et al. (2012) and “Data sources” section of this paper show that the sequence of second-order differences of measured parameters can be used to test the hypothesis of a stable pattern of change. When this hypothesis cannot be rejected, the analyzed data approximate the types of physical conditions that are normally found within a small range of time and space. This outcome was obtained for both parallel and repeated observations that were examined with the Dixon detection scheme. If this short time span can be converted to a window for a continuous time series, the value of this time series window can be set in the range of (3–20) according to the settings in the Dixon lookup table.

Figure 3 is a diagram of the difference analysis, Dixon criterion detection and reconstruction principle.

First, we process the original continuous observation time series using the second-order differences. After the difference processing, we set the value of the window win Dixon detection scheme (the intervals for the segments of the analytical series are \( \{x(i), x(i+w)\} \); the total number of records is \( w \).

Second, we sort this sequence data, select the corresponding calculation formula in (15–18) and calculate the value of \( Q \). After the value of \( Q \) is calculated, we process the outliers according to Table 1 and mark the positions of the outliers in the original sequence according to the sequence information before ordering. Then, we go to the next analytical sequence interval \( \{x(i+1), x(i+w+1)\} \) and repeat the above process.

Finally, in terms of reconstructing the sequence that must be processed to rationalize the outlying data points, we use the median of the data points in the window instead of the outliers; alternatively, we select a common filtering method, e.g., a spline interpolation method, to replace the outliers. In this paper, the filtering method is based on the inverse distance weighted average of the recorded data in the window. Unmarked data are not processed.

Results and comparative analysis

Results

For the measured data, both measured parameters and time data are represented as discrete data sets. For example, the data set of the measured parameters can be expressed as \( \{X_1, X_2, \ldots, X_k, \ldots, X_n\} \), and the time data set is \( \{T_1, T_2, \ldots, T_k, \ldots, T_n\} \). Then, the calculation method for the second-order differences is specified as follows:

\[
\left\{ \begin{array}{l}
F'(X_k, T_k) = \frac{F'(X_{k+1}, T_{k+1}) - F'(X_k, T_k)}{T_{k+1} - T_k} \\
F'(X_k, T_k) = \frac{X_{k+1} - X_k}{T_{k+1} - T_k}
\end{array} \right. \tag{19}
\]

After calculating the second-order differences according to formula (19), we identify the outliers in the sequence as specified according to the Dixon criterion. In this paper, a sequence window of 15 records in length was selected. The significance level \( x \) was defined as 0.01. The values in the sequence were identified as outliers when \( Q > Q_{0.01} \). The inverse distance-weighted average of the values in the window was used to reconstruct the outliers in the series. The time series of second-order differences was used to further elaborate the details of the process of detecting outlying data points. Point \( k \) of the second-order differences in formula (19) is viewed relative to \( k, k+1, \) and \( k+2 \). Specifically, abnormal data points at points \( k, k+1, \) and/or \( k+2 \) may cause abnormal data in the second-order differences at point \( k \). In this paper, tools are used to determine whether points \( k+1 \) and \( k+2 \) are abnormal data.

Figure 4 shows the results of the difference analysis, detection scheme based on the Dixon scheme, and reconstruction method. The method can effectively identify outliers in
Fig. 3. Diagram of the differences, Dixon criterion detection and reconstruction principle. (a) Original water level measurement data, (b) second-order difference sequence, (c1) an interval segment in the difference sequence, (c2) sorting and execution for Dixon detection, (c3) marking of abnormal points and sequence reconstruction.

Fig. 4. Results of the difference analysis, Dixon criterion detection and reconstruction. (a1, a2) RCM-9 data on water level and velocity, (b) GO8050 air pressure data from the barometer, (c) DSSX DO data, (d) MODIS SST data.
measured data at different sampling periods for different measurement techniques. The reconstruction method was based on the identification of outliers and combined with the weighted averages computed in the windows to sample the sequence. Figure 4 shows that the stability of the sequence data was effectively improved. Additionally, the reliability and quality of the data were enhanced. Furthermore, the advantage of this method is that normal data points are not affected. The information in the original sequence has been fully preserved.

Validity evaluation and diagnostic statistic

In statistics, an outlier observation point is distant from other observations (Neter et al. 1996). Cook’s distance or Cook’s D, which was proposed by Cook (1977), is useful for identifying outliers in the observations for predictor variables. To evaluate the effectiveness of the new method in detecting outliers, Cook’s distance is selected in this paper as the key diagnostic statistic value. An observation with Cook’s distance three times larger than the mean Cook’s distance may be an outlier.

Cook’s distance is the scaled change in fitted values. Cook’s distance $D_i$ of observation $i$ is:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j|0})^2}{p \text{MSE}}$$

where $\hat{Y}_j$ is the $j$th fitted response value; $\hat{Y}_{j|0}$ is the $j$th fitted response value where the fit does not include observation $i$; MSE is the mean squared error; and $p$ is the number of coefficients in the regression model. Cook’s distance is algebraically equivalent to the following expression:

$$D_i = \frac{r_i^2}{p \text{MSE} \left(1 - h_{ii}\right)^{2}}$$

where $r_i$ is the $i$th residual, and $h_{ii}$ is the $i$th leverage value.

The fitted regression equation selects the AR model, and the order of the model is determined by the Akaike information criterion (AIC). To evaluate the effectiveness of the new model, the Cook’s distance diagnostic statistic was compared with the results of the new model. The comparative analysis results of the velocity time series data are shown in Fig. 5. The number of outliers diagnosed by Cook’s distance is 50, and the new method detects 75 outliers. The results of the new method contain all outliers of the Cook’s distance method.

The outlier detection results of the remaining four measurement parameter time series are compared with the Cook’s distance diagnostic statistics, as shown in Fig. 6. The comparison results show that the Cook’s distance method detects 27, 4, and 280 outliers in the water level, DO and SST time series, and the new method detects 35, 11, and 372 outliers (which include all points detected by the Cook’s distance method). For the atmospheric pressure measurement time series data, the Cook’s distance method detects 155.
outliers, and the new method detects 142 outliers, so the detection accuracy is 91.61%. The comparative analysis results show that the new method is more effective.

Comparative analysis

The performance of the Dixon criterion method, particularly for continuous time series observations, was examined in this paper. We compared the results of the present study with those reported by Dou, who also used difference data to detect abnormal information. Dou used the $3\delta$ testing method. Figure 7 shows the results of the two methods. The two methods can identify and smooth data sequences with substantial variations in location. The Dou method identifies outliers based on the statistical characteristics of the second-order differences of all analyzed data. In the current study, outliers were identified using statistics based on a sliding window and according to a statistical table for a lookup procedure. The detection theory in the Dou method is a global or all-records procedure. In contrast, in the current study, the detection algorithm is a local test procedure. This contrast is apparent in Fig. 7. The outliers detected by the Dou method appear in locations characterized by substantial variations. In contrast, the methods in the current study can detect both local and global abnormalities in the data.

The comparative analysis in this paper shows that the results of the detection and reconstruction procedures furnished by the $3\delta$ detection method (Dou et al. 2012) can satisfy the data-processing system performance requirements if the global difference data remain within the interval $[\mu-3\delta, \mu+3\delta]$. However, for certain types of continuous time series observation sequence data, effective outlier detection by this method is difficult if the sequence shows only a small anomaly because the measurement series has a long time span and the second-order difference sequence has a complex structure. In contrast, the detection algorithm in the current study applies a sliding-window detection approach, which can satisfy the requirements imposed by the long timespan of the continuous-observation data sequence. Abnormal information can be detected in successive individual stages that form the sequence.

Fig. 6. Comparison of Cook’s distance diagnostic statistic with the results of the new model for the water level, Atm press, DO and SST time series.

Fig. 7. Comparison of the Dixon method and the $3\delta$ method for the velocity second-order differential data (The red line is a plot of the second-order differences of the original data. The green line is a plot of the second-order differences of the $3\delta$ detection dataset and reconstructed data. The blue line is a plot of the second-order differences of the Dixon criterion detection dataset and reconstructed data. Boxes a, b, and c show obvious differences between the two methods).
The quantitative statistical indicators obtained from the evaluated methods for second-order difference data are shown in Table 4. The Max, Min, and SD values portray the level of data quality that is characteristic of a continuous-observation data sequence. The results and various statistical indicators for different methods are compared. The tabulated data show that both methods produce good results, but the statistical indicators for outlier detection according to the Dixon criterion and data reconstruction imply that the Dixon criterion-data reconstruction method yields superior results to the Dou method.

**Discussion and conclusions**

Given the complexity and immense quantity of continuous time series ocean-observation data, effective quality control and a properly designed computer processing algorithm are key data pre-processing steps for data analysis and utilization. This paper presents a new method based on the original Dixon detection criteria. The new method combines two traditional methods: data quality control and Dixon detection theory. The new quality control method builds on the achievements of the Dou method. Accordingly, our method and the Dou method have similarities and differences. The common elements of these two methods are difference analysis and statistical outlier detection. In Dou’s paper, the “data fences,” which are defined by three times the standard deviation of the differences of the time series measurements, are used to define the upper and lower limits to check the existence of data outliers. In contrast, the new method in this paper is based on the original Dixon detection criteria and sliding-window detection technology. The method in Dou’s paper is a quality control method for global data, whereas the new method in this paper is a local quality control method.

Parameter \( n \) is important in Dixon detection theory. The value \( n = 30 \) is a notably frequent choice (Zhang 1995; Shi and Zhou 2011). Other values can also be used (e.g., Fan et al. 2013; Zheng et al. 2014). The sliding window size is closely related to the range of \( n \) values. If we select a larger value of \( n \), the complexity of the calculations increases, but the consequences of outliers in the data can be unexpected. The value of \( w \) is another key parameter of the new method. In this paper, a length of 15 records is used to define the sequence window based on experience and data characteristics. Different values of parameter \( w \) may be used for different observations. In this study, the consequences of different selections of \( w \) values have not been explored. Further research is necessary to find a reasonable and effective method.

The methods in this paper are intended for use with substantial amounts of continuous time series marine observation data. Such data are complex, involve long observation times and are immense in quantity. In this paper, the tests and comparative analyses show that our method has several advantages and disadvantages:

1. It is difficult to conduct parallel repeated observations for marine continuous-observation data. Thus, information mining of data represents an important direction for data quality control. In this paper, a new method is provided to solve this problem based on the Dixon detection criterion, which yields a new tool for the quality control of complex marine observation data. First, a stationarity test based on the ADF method is applied to the observations. The second-order differences of the data satisfy this stationarity test. The physical meaning of this finding is that the measurements show the same physical state over a small range of time and space. This procedure is equivalent to a parallel observation test. Because a small range of temporal and spatial observations corresponds to a short time sequence of data to be addressed by the algorithmic design, this short time sequence is represented by a sliding window. The window size, which is determined by the variations in the observed properties of the data and by researchers who are experienced in addressing the particular characteristics of the data and the issues raised by these local particularities, is an important determinant of the ultimate quality and performance of the algorithm.

2. Compared with the classical Cook’s distance diagnostic statistic outlier detection method, the experimental
results show that the accuracy of this method is satisfactory. The outlier detection is more stringent than the Cook’s distance method.

3. The advantage of the Dixon detection method is that it does not need to estimate the mean or standard deviation of the data subjected to analysis. The method can identify gross errors according to the results of ordering the data for analysis according to the size of the individual observations. The method is realized by calculating the index parameter Q and performing a table lookup in support of the comparative tests. These features of the method improve its computational efficiency.

4. The sliding-window detection technology, which can satisfy the requirements imposed by the long timespan of continuous time series observation data sequences and the complexity of second-order difference sequences, is used in the new algorithm. This feature of the new algorithm contrasts with the features of the standard reference Dou method. The outliers in the successive stages of a data sequence can be easily detected by our method.

Note that the outliers (e.g., minimum and maximum values) detected based on the Dixon criterion are identified using the window-sequence sorting method. Other anomalous points in the sequence cannot be detected. Accordingly, this criterion cannot detect multiple abnormal points in a given window. The sliding-window technology can overcome this shortcoming. However, a concrete analysis of the choice of a time series window remains necessary. If the outliers in the sequence fall outside the selected window, it may be difficult to efficiently detect them using this method.

References

Alvera-Azcárate, A., A. Barth, and M. Rixen. 2005. Reconstruction of incomplete oceanographic data sets using Empirical Orthogonal Functions. Application to the Adriatic Sea. Ocean Model. 9: 325–346. doi:10.1016/j.ocemod.2004.08.001

Beckers, J.-M., and M. Rixen. 2003. EOF calculations and data filling from incomplete oceanographic data sets. J. Atmos. Ocean. Technol. 20: 1839–1856. doi:10.1175/1520-0426(2003)20<1839:EOCFCA>2.0.CO;2

Chen, S. 1991. Marine data processing analysis method and algorithm, 1st ed. China Ocean Press.

Cook, R. D. 1977. Detection of influential observations in linear regression. Technometrics Am. Stat. Assoc. 19: 15–18. doi:10.2307/1268249

Cosme, E., J.-M. Brankart, J. Verron, and P. Brasseur. 2010. Implementation of a reduced rank square-root smoother for high resolution ocean data assimilation. Ocean Model. 33: 87–100. doi:10.1016/j.ocemod.2009.12.004

Dickey, D. A., and W. A. Fuller. 1979. Distribution of the estimators for autoregressive time series with a unit root. J. Am. Stat. Assoc. 74: 427–431. doi:10.1080/01621459.1979.10482531

Doong, D. J., S. H. Chen, C. C. Kao, B. C. Lee, and S. P. Yeh. 2007. Data quality check procedures of an operational coastal ocean monitoring network. Ocean Eng. 34: 234–246. doi:10.1016/j.oceaneng.2006.01.011

Dou, W., J. Jiang, and B. Zhou. 2012. An algorithm for the difference correlation filter for multi-parameter marine timing observation data. Acta Ocean. Sin. 34: 51–58.

Fan, D., W. Zhu, and Y. Pan. 2013. Noise detection for NDVI time series based on Dixon’s test and application in data reconstruction. J. Remote Sens. 17: 1158–1174. doi:10.11834/jrs.20132274

Greene, W. H. 2003. Econometric analysis, 5th ed. Prentice Hall.

Hamilton, J. D. 1994. Time series analysis. Princeton Univ. Press.

Huang, L., S. Zhang, and W. He. 1993. A new quality control method for ocean hydrological data. J. Beijing Polytech. Univ. 19: 92–97.

Huang, M., G. Zhai, and R. Wang. 1999. The detection of abnormal data in marine survey. Acta Geod. Cartograph. Sin. 28: 269–277. doi:10.3321/j.issn:1001-1595.1999.03.015

Ingleby, B., and M. Huddleston. 2007. Quality control of ocean temperature and salinity profiles-Historical and real-time data. J. Mar. Syst. 65: 158–175. doi:10.1016/j.jmarsys.2005.11.019

Ji, F., C. Miao, and S. Zhang. 2004. Quality control of Argo data based on climatological T-S relationship. Mar. Sci. Bull. 23: 8–15. doi:10.3969/j.issn.1001-6392.2004.06.002

Lei, H. 1997. Comparison and discussion of gross error detection methods. Pet. Instrum. 11: 54–64.

Li, B. 2012. Design of data buoy receiving system based on various communication technology and data analysis. Master’s thesis. Ocean Univ. of China.

Ma, Y., S. Xu, and S. Li. 1999. Data set of the maritime climate and its quality control. Meteorol. Month. 25: 15–18. doi:10.3969/j.issn.1000-0526.1999.02.003

McKinney, J. P. 1993. A method for locating spikes in a measured times series, p. 388–393. Proceedings of the second Symposium on Ocean Wave Measurement and Analysis, New Orleans, LO, USA.

Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 1996. Applied linear statistical models, 4th ed. Irwin.

Pan, D., T. Li, and Y. Bai. 2012. Ocean: Earth the greatest carbon library. Ocean. Res. 28: 92–97. doi:10.3969/j.issn.1001-6392.2014.06.002

Sha, D. 2003. Error analysis and uncertainty evaluation of measurement. China Metrology Press.

Shi, J. 2010. The data quality control method research and software realization for marine environment real-time observation. Master’s thesis. National Ocean Technology Center.

Shi, J., and Z. Zhou. 2011. Study on technology of ocean station data quality control. Ocean Technol. 30: 114–117. doi:10.3969/j.issn.1003-2029.2011.01.028
Shi, M., G. Gao, and X. Bao. 2000. Marine survey method, 1st ed. Qingdao Ocean Univ. Press.

Wang, H., R. Zhang, and G. Wang. 2012. Quality control of Argo temperature and salinity observation profiles. Chin. J. Geophys. **55**: 577–588. doi:10.6038/j.issn.0001-5733.2012.02.020

Wong, A. P. S., G. C. Johnson, and W. B. Owens. 2003. Delayed-Mode calibration of autonomous CTD profiling float salinity data by β-S climatology. J. Atmos. Ocean. Technol. **20**: 308–318. doi:10.1175/1520-0426(2003)020<0308:DMCOAC>2.0.CO;2

Zhang, Y. 1995. Examination procedure of Dixon method in application to environmental monitoring. J. Shanghai Environ. Sci. **14**: 38. doi:10.11840/j.issn.1001-6392.2014.02.017

Zheng, L., Y. Liu, and W. Cui. 2014. Research on the quality assessment of marine environmental monitoring data. Mar. Sci. Bull. **33**: 228–234.

Zhu, H. 2009. Intelligent voyage ocean monitoring system. Master’s thesis. Ocean Univ. of China.

**Acknowledgments**

This research was supported by the National Natural Science Foundation of China (Nos. 41401404), the National One-Thousand-Talent Scheme Foundation and the Open Research Fund of State Key Laboratory of Estuarine and Coastal Research (Nos. SKLEC-KF201613).

**Conflict of Interest**

None declared.

*Submitted 26 October 2016  
Revised 05 June 2017  
Accepted 05 September 2017*  

*Associate editor: Gregory Cutter*