Journal of Oceanology and Limnology
Vol. 37 No. 6, P. 1983-1993, 2019
**https://doi.org/10.1007/s00343-019-8258-y**

# Quality control of marine big data—a case study of real-time observation station data in Qingdao*

QIAN Chengcheng[1, 3], LIU Aichao[1, **], HUANG Rui[1], LIU Qingrong[1], XU Wenkun[2], ZHONG Shan[1], YU Le[1]

[1] *North China Sea Marine Forecasting Center of State Oceanic Administration, Qingdao 266061, China*

[2] *Qingdao Geotechnical Investigation and Surveying Research Institute, Qingdao 266000, China*

[3] *Laboratory for Regional Oceanography and Numerical Modeling, Qingdao National Laboratory for Marine Science and Technology, Qingdao 266000, China*

Received Sep. 21, 2018; accepted in principle Dec. 17, 2018; accepted for publication Jan. 23, 2019


**Abstract** Offshore waters provide resources for human beings, while on the other hand, threaten them because of marine disasters. Ocean stations are part of offshore observation networks, and the quality of their data is of great significance for exploiting and protecting the ocean. We used hourly mean wave height, temperature, and pressure real-time observation data taken in the Xiaomaidao station (in Qingdao, China) from June 1, 2017, to May 31, 2018, to explore the data quality using eight quality control methods, and to discriminate the most effective method for Xiaomaidao station. After using the eight quality control methods, the percentages of the mean wave height, temperature, and pressure data that passed the tests were 89.6%, 88.3%, and 98.6%, respectively. With the marine disaster (wave alarm report) data, the values failed in the test mainly due to the influence of aging observation equipment and missing data transmissions. The mean wave height is often affected by dynamic marine disasters, so the continuity test method is not effective. The correlation test with other related parameters would be more useful for the mean wave height.

**Keyword**: quality control; real-time station data; marine big data; Xiaomaidao Station; marine disaster

## 1 INTRODUCTION

Development and utilization of marine resources by humans are mainly concentrated in coastal waters. More than half of the world's population lives within 100 km of a coast, and 40% of the population in China lives in 11 coastal provinces. Seventy percent of fish resources and almost all of the marine oil and gas resources are concentrated in offshore areas (Shi et al., 2008). Offshore areas have significant meaning for the development of human beings. The offshore areas provide basic resources for the survival of human beings, and at the same time, marine disasters can harm human beings. Marine disasters can be divided into dynamic disasters and ecological disasters. The former is mainly composed of the extreme dynamic marine events, such as tsunamis, marine waves, and storm surge. Ecological disasters are mainly caused by human activities, such as red tides, green tides, and oil spills. According to the China Marine Disasters Bulletin (SOA, 2018), marine disasters directly resulted in economic losses of 6.4 million Yuan in 2017, which did not consider a heavy disaster year. Therefore, the reasonable exploitation, utilization, and protection of the oceans require real-time ocean observations.

Ocean stations are observation facilities that are built on coastal beaches and house observation instruments that support the long-term observation of offshore marine hydrometeorological factors. China was one of the first countries to develop marine observation stations in the world. As early as 1905, construction began on the tide station in Qingdao, which opened the door for ocean observations (North China Sea Branch of the State Oceanic Administration,

* Supported by the National Key Research and Development Program of China (Nos. 2016YFC1402000, 2018YFC1407003, 2017YFC1405300)
** Corresponding author: liuaichao@ncs.mnr.gov.cn

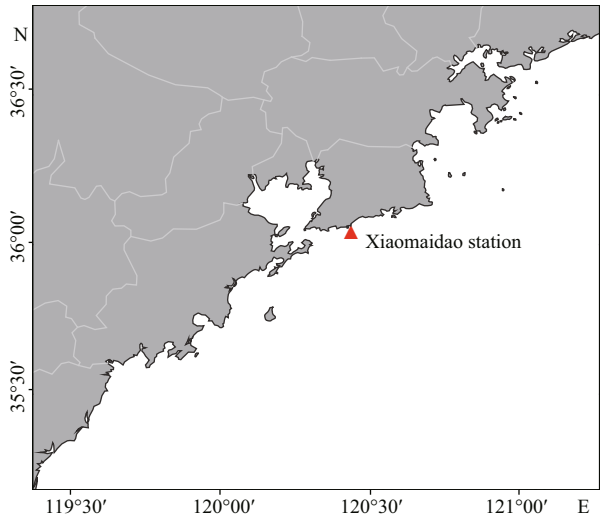


**Fig.1 The location of the Xiaomaidao station**

1993). There were 524 coastal stations all together in China in 1997, including 61 ocean stations, 191 tide stations, 113 meteorological stations, 158 seismic stations, and 1 radar station. At present, China has built approximately 150 wave, thermohaline, and meteorological element observation stations on the coast, and there are 54 stations in the North China Sea area. Ocean station observation systems pave the way for the extensive tide, wave, temperature, salt, sea ice, weather, and pollution projects, such as observation and surveillance projects.

However, due to the location, equipment, maintenance, and communication, the data detected by these stations can have abnormal values. At the same time, the changes to the ocean station observation data over time are quite significant and complex, and some changes do not exist alone but will interact, thus increasing the significance of the change in observed elements over time (Thadathil et al., 1998; Ingleby and Huddleston, 2007). Therefore, before analyzing ocean station observation data, necessary quality control must be performed. To detect data errors, a prompt method is used to modify and verify suspicious or abnormal data, thus reducing the errors in the data application. This procedure is conducted to ensure the reliability, representativeness, and comparability of the data and reduce the adverse effects of poor data quality on government decision making.

The traditional real-time data quality control method is manual control. With the continuous development of science and technology and the increase in the amount of real-time data, marine observation data are being considered marine big data. The gradual replacement of humans by artificial intelligence or computers has become an inevitable trend (NOAA and Integrated Ocean Observing System (IOOS) Program Office, 2008). Different data quality control methods have different limitations. Reasonable quality control methods should be based on different occasions and elements. The date, location, format, and inspection information are often used for widely used data quality control methods such as basic observation range tests, statistical and climate characteristic tests, continuity tests, and correlation tests.

Currently, the automated quality control of data in meteorological offices and stations is sufficient, but the quality control procedures for ocean station observation data are still in development. Many publications have studied the quality control of observation data, but most have focused on ground meteorological data. Marine data research is rarely reported, and only a few people have used delayed low-resolution observation data for quality control of marine data (Li and Li, 1997; Yang et al., 2017). This paper uses hourly real-time ocean station observation data to discuss quality control techniques and methods to find the most suitable quality control method for the data collected at the Xiaomaidao station. The second section introduces the real-time data from the station, the third section introduces the method, the fourth section presents the results and discussion, and the last section summarizes the study.

## 2 DATA

This study used hourly real-time observation data from the Xiaomaidao station from June 1, 2017, to May 31, 2018, and the parameters included mean wave height, air temperature, and pressure. The Xiaomaidao environment monitoring station was built in 1953 and is located on Xiaomaidao Island, Laoshan District, Qingdao, 120.4258°E, 36.0526°N, as shown in Fig.1. The long-term, continuous, and dynamic monitoring of the Qingdao coastal waters provides access to a large number of representative hydrological and meteorological observation data that can be utilized for marine disaster prevention and mitigation, economic construction, transportation, and scientific research, and these data play an important role in national defense construction.

The standard file structure of the observation data from the Xiaomaidao station consists of three parts: data title, data record, and data description. Hydrological measurements are recorded on an hourly basis. Hourly meteorological elements are observed as well by the marine monitoring station,

including temperature, pressure, relative humidity, precipitation, wind speed, and wind direction. The receiving and communicating data are in data file format and use ASCII code. There is space between each data point. The storage formats and examples of wave, air temperature, and pressure data are as follows. In addition, they are the standard format of the marine station observation data.

(1) Air temperature

The filename: atMMDD IIIII

The file content and format are as follows:

The temperature data file is only one row, with 24 hourly values and two extreme values.

YYYYMMDD + + < space >< 21 points measured value > + < space > + < 22 points measured value > + < space > + < 23 > + < space > + < 00 PM measurements > + < space > + < 01 point measurements > +...+ + < space ><> 20 points measured + < space > + < day high > + < space > + minimum <, >, < enter > a new line

(2) Pressure

The filename: bpMMDD IIIII

The file content and format are as follows:

The air pressure data file is only one row, with 24 hourly values and two extreme values.

YYYYMMDD + + < space >< 21 points measured value > + < space > + < 22 points measured value > + < space > + < 23 > + < space > + < 00 PM measurements > + < space > + < 01 point measurements > +...+ + < space ><> 20 points measured + < space > + < day high > + < space > + minimum <, >, < enter > a new line

(3) Wave

The filename: wvMMDDHH IIIII

The file content and format are as follows:

< YYYYMMDDHHMM > + < wave sampling interval > + < mean wave height > + < average period > + < the maximum wave >, < the maximum cycle > + < one tenth wave height > + < one tenth period > + < a third wave height > + <a third period> + < wave number > + < wave direction > + < enter >

## 3 METHOD

The following three principles function as guidance for the quality control of the real-time observation data from the ocean station (Kearns et al., 2004):

(1) The format of the data files conforms to the standard format requirements.

(2) The observed values are consistent with the physical characteristics of the marine environment.

(3) The observations comply with the spatial and temporal characteristics of the element.

This study uses the following eight methods for the hourly real-time mean wave height, air temperature,and pressure observations data on the basis of the above three principles (National Data Buoy Center, 2009; Wan Daud, 2010; Morello et al., 2011; Xu et al., 2014).

(1) The date of the test

The observation date should be within a reasonable scope. The year values should not be greater than the current year. The month value should be in range of 1–12. The date values should be between the number of days during the month. The hour values should be in range of 0–23, and the minutes and seconds values should in range of 0–59.

(2) Location test

The location of the ocean observation station should be within a reasonable range; for example, the latitude should be from -90° to 90°, the longitude should be from -180° to 180°, and the fixed observation position drift range should be less than 5 km based on the technical specification for quality control of marine observation delay data.

(3) Format test

The marine observation data should be in accordance with the prescribed format; for example, the project elements used to record the starting location and length, the data record types, and the missing values should meet the corresponding requirements (Yu et al., 2010). The data should be checked according to the fixed format.

(4) Unique value test

Some elements in observation records remain unchanged for long periods, and some of these records can have unique values, such as data types, buoys, platform codes, marine observation stations, observation methods, instrument names, observation instrument altitudes, and observation point depth. The values and conventions of these records must be consistent.

(5) Range test

The statistical analysis of the existing domestic and international marine observation data indicates that many of the elements have different characteristics. If the data are beyond the range that is considered normal, they could be abnormal. Generally, the values of the elements should be within the range of the extreme values from previous years.

(6) Continuity test

The marine observation elements should be continuous within a certain time and space. This

condition means that the differences in the observation values between two adjacent times or locations should be within a certain range. The specific inspection method is as follows: assuming the current observation value of $v(t)$ and adjacent value of $v(t–1)$, test value=$|v(t)–v(t–1)|$.

(7) Statistical and climate characteristics test

In the theory of marine observation, data often have certain probabilities and statistical properties that correspond to random variables and random processes. The data should be independent and obey a specific distribution. The time series data should correspond to a random process and should also be stable or cyclical. Independent values are often mistakes. According to the variations of the marine environment, it should be evaluated whether the data follow the characteristics of different time scales, such as seasonal and daily variations. In this paper, the periodic and the standard deviation are used to perform the quality control.

(8) Correlation test

The relationships between elements in marine observation data, including autocorrelation and other correlations, should be evaluated.

For the autocorrelation evaluation, each time record or hourly value on one day should be beyond the extreme value, the maximum wave height value should be larger than the mean wave height, and the maximum period must be greater than the mean period.

Other correlations include the relationship among the wave type, wave height, and sea condition, the relationship among wind speed, wave height, and wave period, and the relationship among salinity, temperature, and density.

## 4 RESULT AND DISCUSSION

According to the eight methods in the Section 3, we here evaluate the hourly mean wave height, air temperature, and pressure data from June 1, 2017, to May 31, 2018, for a total of 8 736×3 data records at the Xiaomaidao station. After the dates, locations, formats, and unique values are evaluated, there are 103 groups of mean wave heights that have only data title records and lack the other data records. During the following quality control check, the missing data are given the default value of 999. All of the temperature and pressure data pass the location, format and unique value tests in the first round of data quality control. The original mean wave height, temperature, and pressure data are shown in Fig.2. Figure 2 shows that there are many default values in the original data, and we eliminate the default values and record the default number as Error1. In the 8 736×3 groups of data, there are 191 default wave height values (including the first round test to determine unqualified data), there are 121 default temperature and pressure values. The fractions of defective wave height, temperature, and pressure data are 2.2%, 1.4%, and 1.4%, respectively, and the processed data are shown in Fig.3.

**Table 1 The numbers of the three types of errors and the standard deviation of the data that passed the tests**

| Parameter | Error1 | Error2 | Error3 | Standard deviation |
|---|---|---|---|---|
| Mean wave height | 191 | 90 | 615 | 0.21 |
| Air temperature | 121 | 51 | 849 | 9.93 |
| Pressure | 121 | 0 | 0 | 8.87 |

For the range test, the standard value should be the extreme value from many years according to the historical data and the experience of the forecasters from the Xiaomaidao station, and the wave height value should not be more than 4 m. In this paper, the air temperature and pressure standards are based on the effective monitoring range of the observation instrument. The temperature is between -35 and 45°C, and the pressure is between 800 and 1 100 hPa. According to these ranges, 90 mean wave height data records do not pass the test, 51 temperature data records do not pass the test, and all of the pressure data records pass the test. The data that do not pass the test are recorded as Error2. The fractions of defective mean wave height, temperature, and pressure data records are 1.1%, 0.6%, and 0%, respectively, and the processed data are shown in Fig.4.

According to historical data, aside from special sea conditions, the mean wave height difference at the Xiaomaidao station between each adjacent hour is no greater than 2 m, and the temperature difference between adjacent hours is no greater than 4°C. According to the standard of the mean wave height and temperature continuity test, it was found that 615 groups of mean wave height data did not pass the test, and 849 groups of temperature data did not pass the test. The data that did not pass the test were recorded as Error3. The fractions of defective mean wave height and temperature data are 7.3% and 9.9%, respectively. The processed data are shown in Fig.5.

In the statistical and climate characteristics tests, the standard deviations of the mean wave height, air temperature, and pressure are calculated as 0.21, 9.93, and 8.87, respectively (Table 1). The time series of the three standardized parameters are shown in Fig.6.

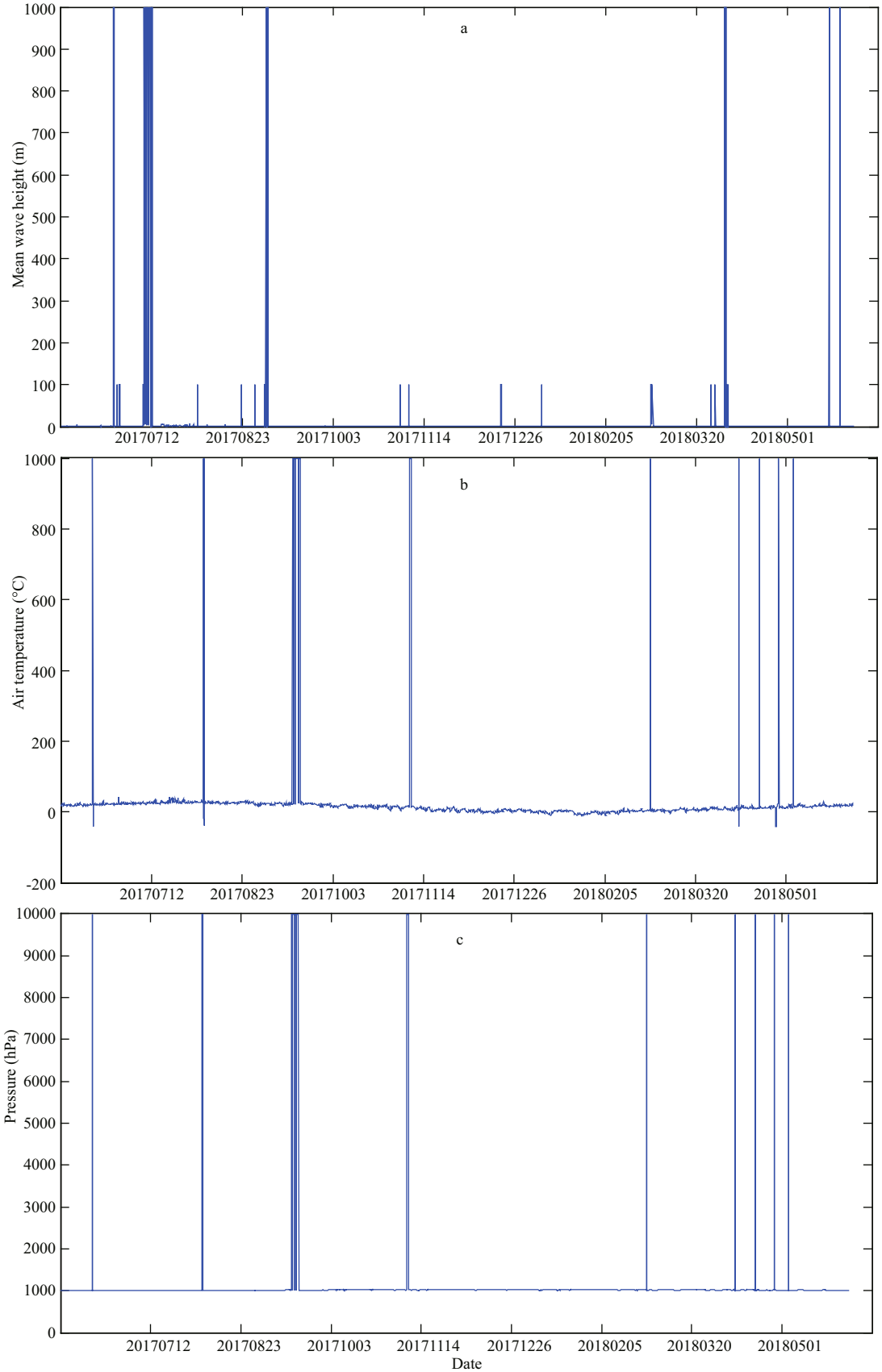


**Fig.2 The time series of original real-time mean wave height (a), air temperature (b), and pressure (c) data**

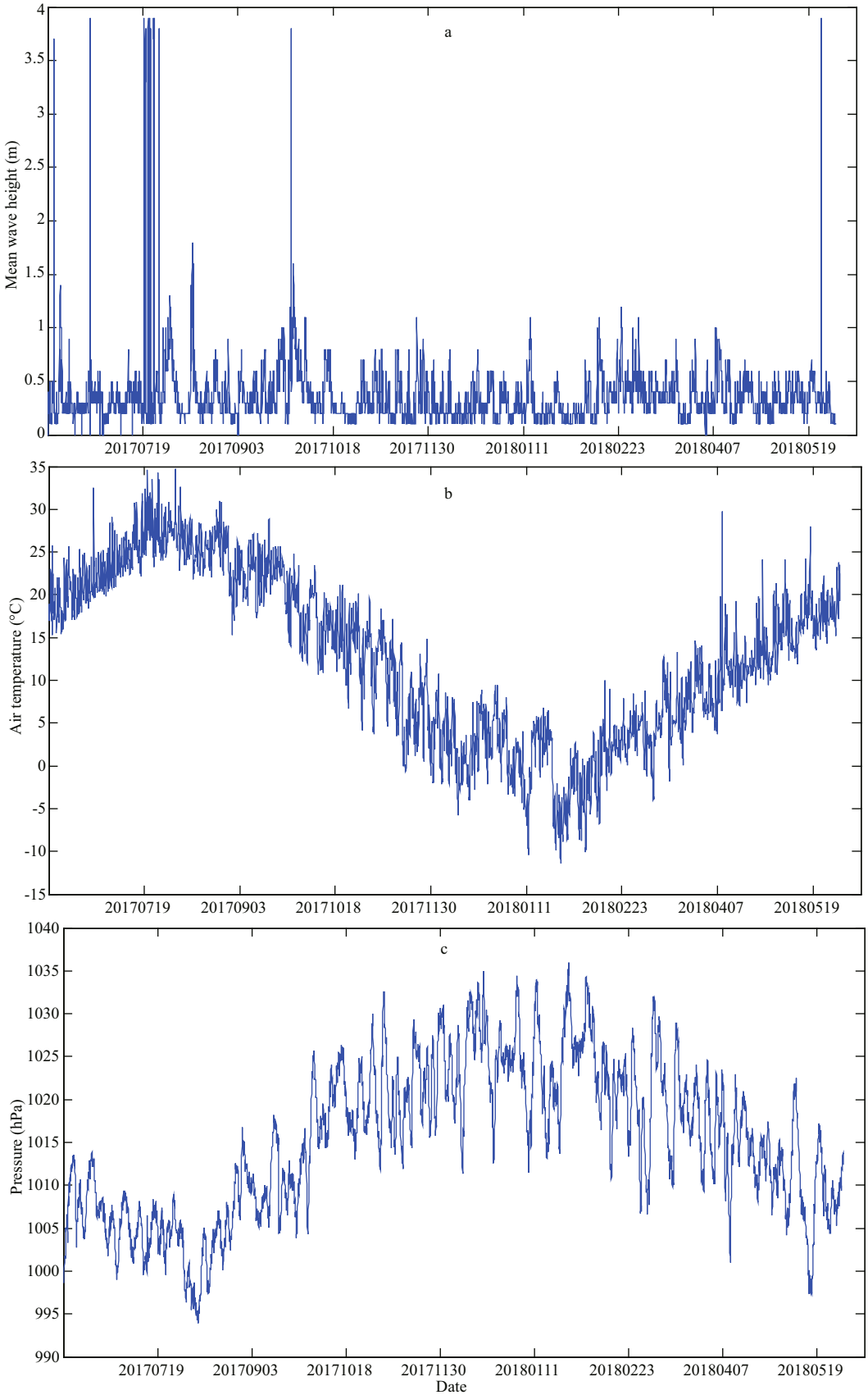


**Fig.3 The time series of real-time mean wave height (a), air temperature (b), and pressure (c) data without missing values**

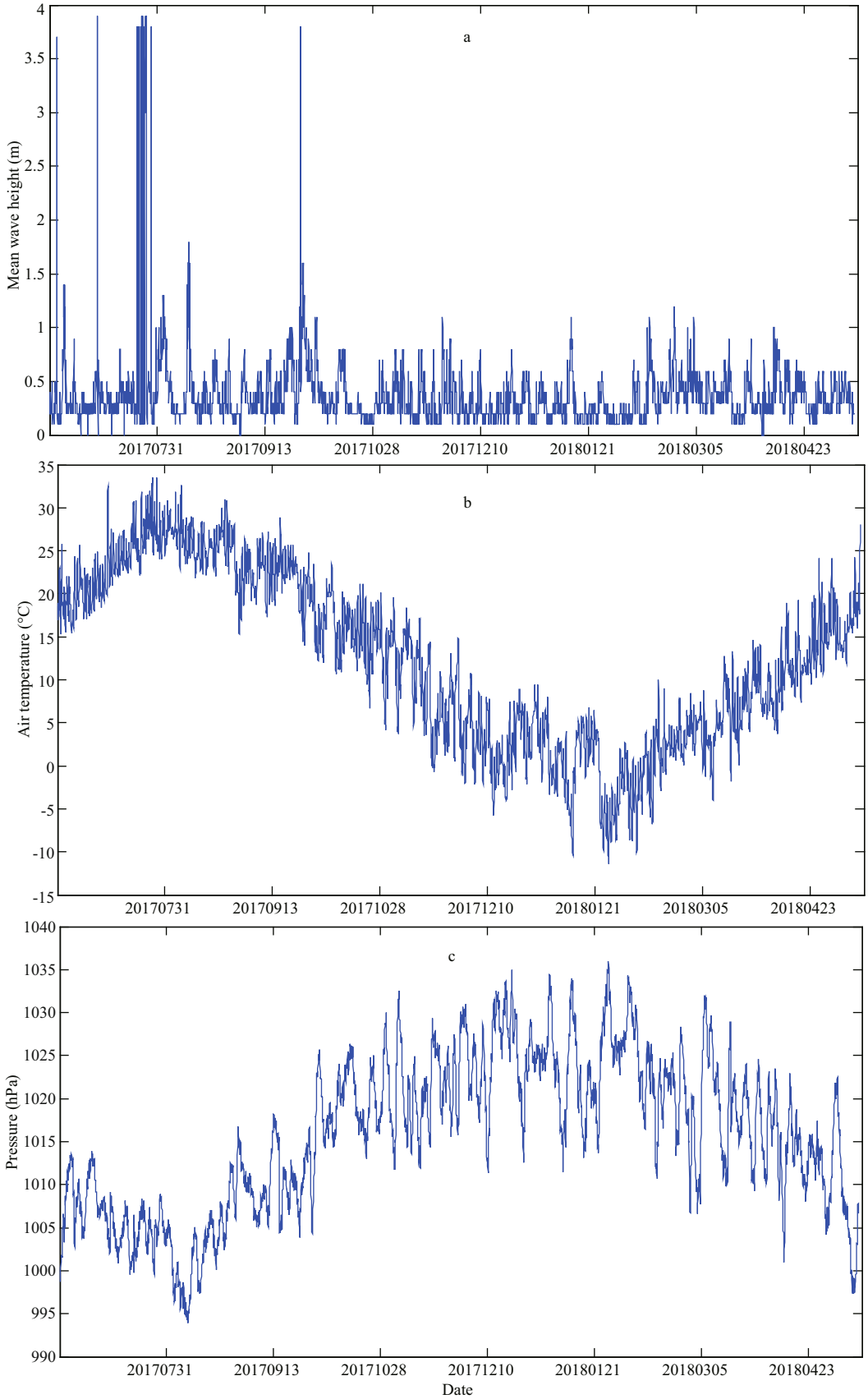


**Fig.4 The time series of real-time mean wave height (a), air temperature (b), and pressure (c) data that passed the range test**

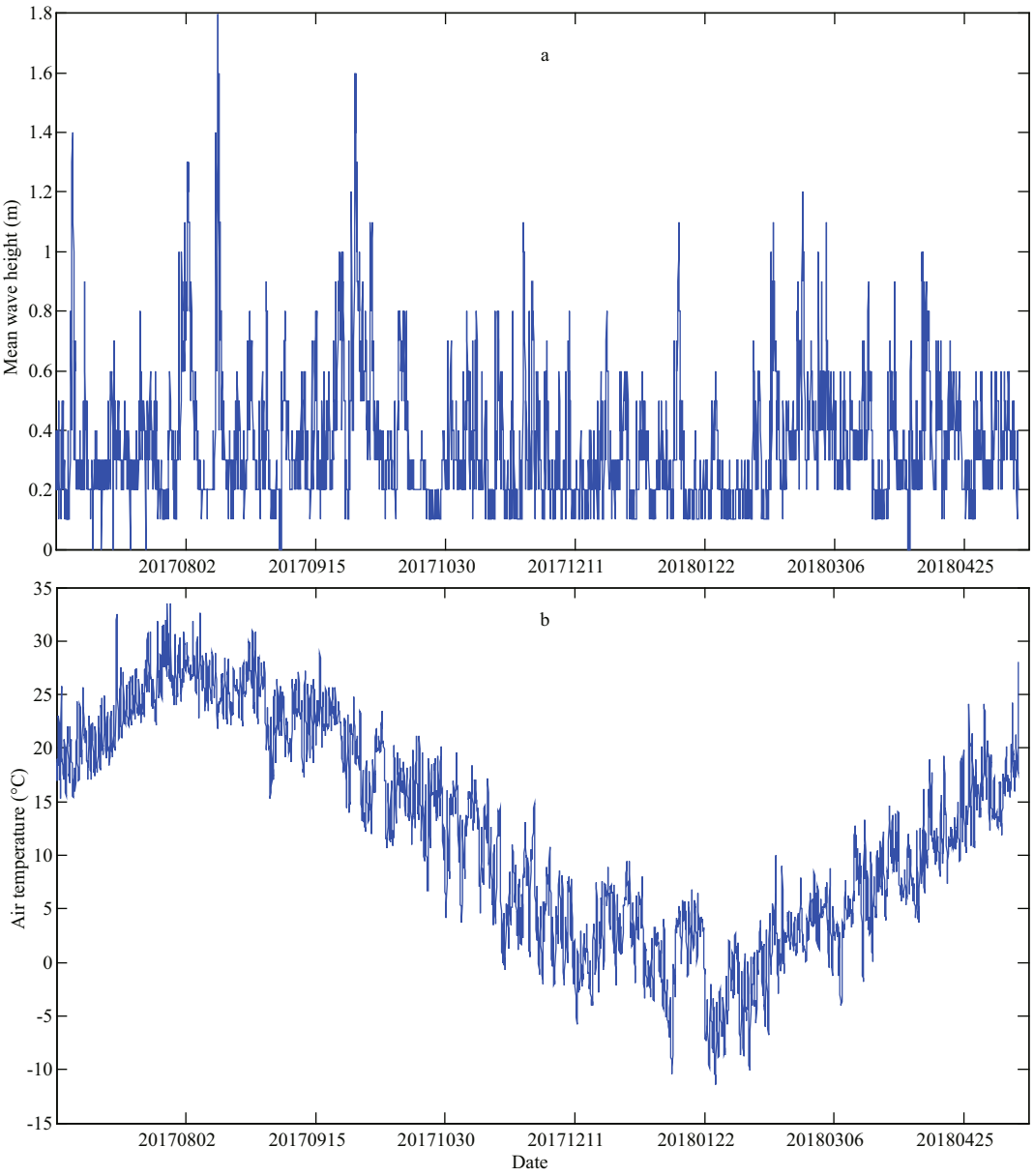


**Fig.5 The time series of real-time mean wave height (a) and air temperature (b) data that passed the continuity test**

The pressure is the same as Fig.4c.

The diurnal, seasonal, and annual variations of the air temperature and pressure are obvious, and the diurnal variation of the mean wave height is featured, but the seasonal and annual changes are not significant.

The correlation coefficients for these three parameters were calculated. The correlation coefficient between the mean wave height and air temperature is 0.18, the correlation coefficient between the mean wave height and pressure is 0.21, and the correlation coefficient between air temperature and pressure is -0.82. These results indicate that temperature and pressure have obvious negative correlations and can be directly compared using the correlation test. From Fig.6, after the range test and continuity test, the air temperature and pressure data are reasonable, but the mean wave height data still has some abnormal values. The correlation test seems more necessary for the mean wave height data.

A total of 7 427 groups of air temperature and pressure data passed the previous quality control checks. A scatter plot of the air temperature and pressure data is drawn; the sum of the sine fitting is shown in Fig.7a, and the density diagram is shown in Fig.7b. After the calculations, the root mean square error (RMSE) is 4.9, and the adjusted $R$ square is 0.69. This test reveals that the temperature and

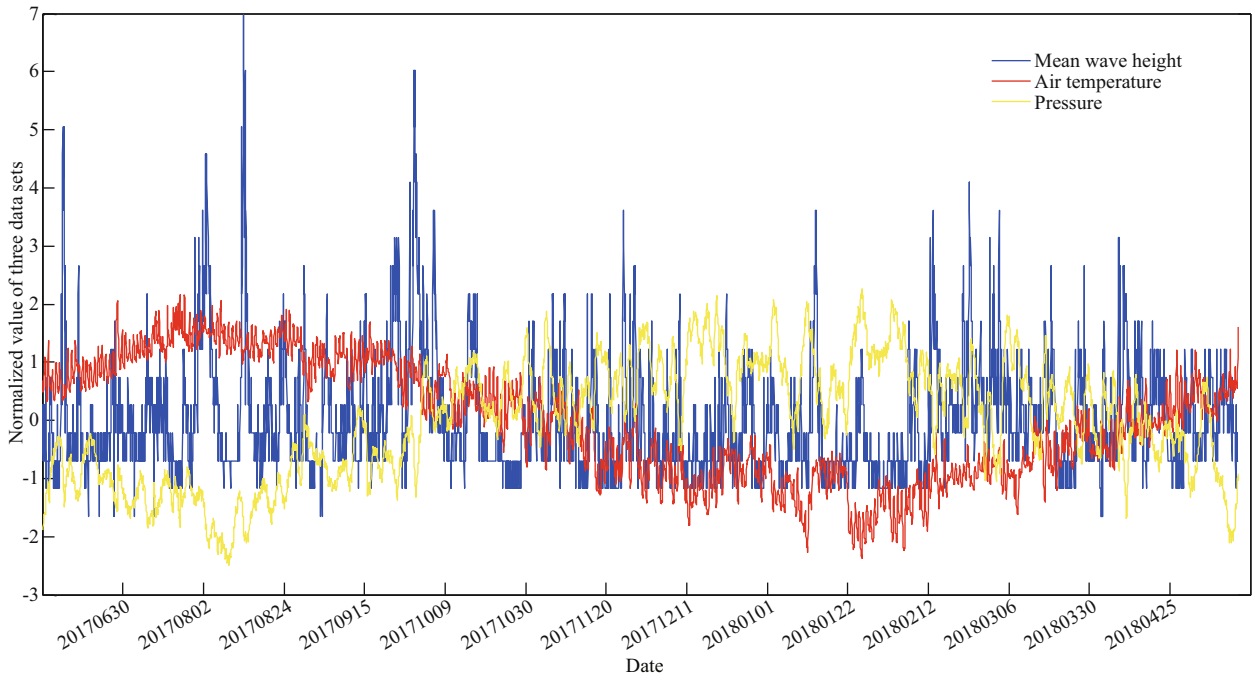


**Fig.6 The standardization time series of three data sets that passed the quality control tests**

Blue line: the mean wave height; red line: the air temperature; yellow line: the pressure.

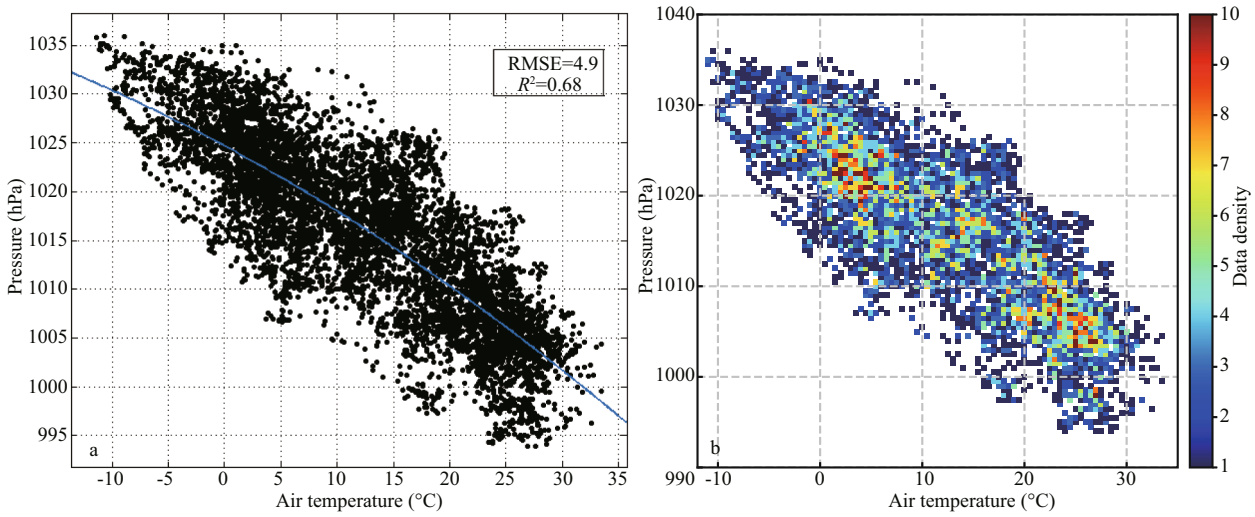


**Fig.7 The scatter plots of the collocated air temperature versus pressure datasets from June 1, 2017, to May 31, 2010**

a. the curved fitting is plotted as the blue line; b. the data density plotted on a logarithmic scale.

pressure data have no abnormal values. The correlation test of the mean wave height is mainly based on the alarm reports of ocean waves. Because the Xiaomaidao station has a special geographic location, its high mean wave height is often lower than the conventional definition of an alert value (more than 2.5 m); therefore, the standard alarm value for the Xiaomaidao station uses a mean wave height of 1 m, according to historical records. A comparison of real alarm report data and mean wave height values that are greater than 1 m is shown in Table 2. As shown in Table 2, there are nine alarm reports, but 10 data records have mean wave heights greater than 1 m. Among these data, three of the times overlap (2017.08.13, 2017.09.30, and 2018.03.04), the mean wave height is close to 1 m at 2 times (2017.11.17 and 2018.04.03), the mean wave height is unusually low 4 times (2017.10.09, 2017.10.28, 2018.01.22, and 2018.03.15) and the mean wave height is abnormally high 7 times (2017.06.06, 2017.08.02, 2017.10.06, 2017.11.25, 2018.01.14, 2018.02.13, and 2018.02.24). After the analysis of the 7 times with abnormally high values, it

**Table 2 Comparison between alarm report data and the mean wave height data with values greater than 1 m**

| Alarm report data | Mean wave height data | Mean wave height |
|---|---|---|
| | 2017.06.06 | 1.2 |
| | 2017.08.02–08.03 | 1.1 |
| 2017.08.13–08.14 | 2017.08.13–08.14 | 1.1 |
| 2017.09.30–10.03 | 2017.09.30–10.03 | 1.1 |
| | 2017.10.06 | 1.1 |
| 2017.10.09–10.10 | | 0.3 |
| 2017.10.28–10.29 | | 0.3 |
| 2017.11.17 | | 0.8 |
| | 2017.11.25 | 1.1 |
| | 2018.01.14 | 1.1 |
| 2018.01.22 | | 0.2 |
| | 2018.02.13 | 1.1 |
| | 2018.02.24 | 1.1 |
| 2018.03.04–03.05 | 2018.03.04–03.05 | 1.1 |
| 2018.03.15 | | 0.4 |
| 2018.04.03 | | 0.7 |

was found that 3 of these times were truly abnormal with hourly mean wave heights greater than 1 m (2017.10.06, 2017.11.25, and 2018.02.13), and 4 of these times were the result of procedural anomalies, and these should be checked with other methods. Therefore, the correlation analysis reveals that the mean wave height has 7 abnormal values (4 abnormally low values and 3 abnormally high values).

We do the histogram of error data to identify the result of the QC. As shown in Fig.8, the "outliers" follows a normal distribution. The result indicates that the QC steps are more reliable and credible. We will do the further study to employs the Bayesian probability theory to take into better account factors such as the accuracy of the reference field itself, so the 'outliers' may be dealt with more carefully (Lorenc and Hammon, 1988; Xu and Ignatov, 2014).

## 5 CONCLUSION

After the eight quality control steps, the qualification ratios of the mean wave height, air temperature, and pressure data were 89.6%, 88.3%, and 98.6%, respectively. In addition, the effective methods to control the air temperature and pressure are range test and continuity test. The mean wave height is often influenced by dynamic marine disasters, so the continuity test method is not applicable. The quality control of mean wave height

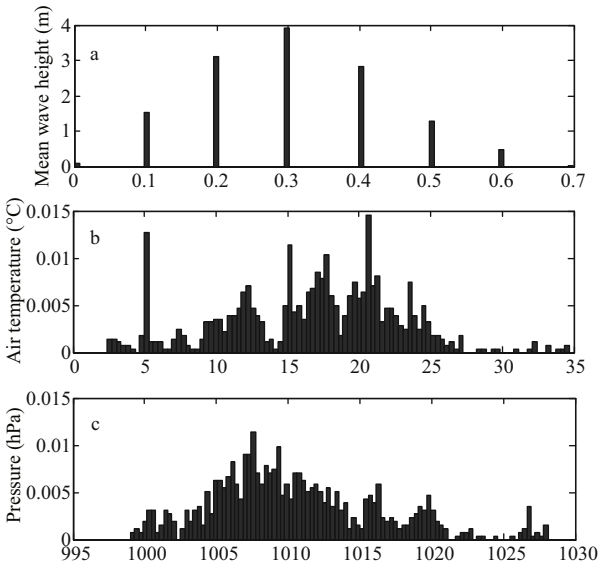


**Fig.8 Histogram of error data**

a. mean wave height; b. air temperature; c. pressure.

data should focus on correlation test methods. The significant errors in the three parameters are caused by the aging of the observation equipment and missing data transmissions. Therefore, in daily work, we should focus on checking the stability of observation equipment and data transmission. The other observation parameters should be studies in the future research.

## 6 DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

**References**

Ingleby B, Huddleston M. 2007. Quality control of ocean temperature and salinity profiles—historical and real-time data. *J. Marine Syst*., **65**(1-4): 158-175, https://doi.org/10.1016/j.jmarsys.2005.11.019.

Kearns E, Woody C, Bushnell M. 2004. QARTOD-I Report. First Workshop Report on the Quality Assurance of Real-Time Ocean Data. December 3-5, 2003. National Data Buoy Center, NWS/NOAA, Stennis Space Center, MS. 89pp, https://doi.org/10.25607/OBP-380. Accessed on 2018-04-23.

Li X K, Li F J. 1997. Marine hydro-meteorological real-time data quality control. *Mar. Forecasts*, 1**4**(3): 71-79. (in Chinese)

Lorenc A C, Hammon O. 1988. Objective quality control of observations using Bayesian methods: theory, and a practical implementation. *Quart. J. Roy. Meteor. Soc*., **114**(480): 515-543, https://doi.org/10.1002/qj.49711448012.

Morello E B, Lynch T P, Slawinski D, Howell B, Hughes D,

Timms G P. 2011. Quantitative quality control (QC) procedures for the Australian national reference stations: sensor data. *In*: Proceedings of Oceans'11MTS/IEEE KONA. IEEE, Waikoloa, Hawaii, USA.

National Data Buoy Center.2009. Handbook of Automated Data Quality Control Checks and Procedures. Stennis Space Center, Mississippi, USA.

NOAA, Integrated Ocean Observing System (IOOS) Program Office. 2008. Data Integration Framework (DIF) Customer Implementation Project Summary and Performance Assessment Plan, Version 1.1. NOAA, IOOS, Quebec City, QC, Canada.

North China Sea Branch of the State Oceanic Administration. 1993. The North China Sea Marine Hydrology and Climate. Ocean Publishing House, Beijing. p.63-190. (in Chinese)

Shi M C, Gao G P, Bao X W. 2008. Methods of Marine Survey. China Ocean University Press, Qingdao, China. p.6-123. (in Chinese)

SOA (State Oceanic Administration, China). 2018. China Marine Disasters Bulletin. http://gc.mnr.gov.cn/201806/t20180619_1798021. html.Accessed on 2018-04-23. (in Chinese)

Thadathil P, Ghosh A K, Pattanaik J, Ratnakaran L. 1998. A quality-control procedure for surface temperature and surface layer inversion in the XBT data archive from the Indian Ocean. *J*. *Atomos*. *Ocean Technol*., **16**(7): 980-982, https://doi.org/10.1175/1520-0426(1999)016<0980:AQCPFS>2.0.CO;2.

Wan Daud W M N. 2010. Quality control for unmanned meteorological stations in Malaysian meteorological department, https://www.wmo.int/pages/prog/www/IMOP/publications/IOM-109_TECO-2012/Session2/P2_01_WanDaud_QC_Unmanned_Meteorological_Stations.pdf. Accessed on 2018-04-23.

Xu F, Ignatov A. 2014. In situ SST quality monitor (*i*Quam). *J*. *Atomos*. *Ocean*. *Technol*., **31**(1): 164-180, https://doi.org/10.1175/JTECH-D-13-00121.1.

Xu J, Yu D T, Yuan Z J, Li B, XuZ Z. 2014. Implementation of marine environment monitoring data quality control system. *Adv*. *Mater*. *Res*., **926-930**: 4 254-4 257, https://doi.org/10.4028/www.scientific.net/AMR.926-930.4254.

Yang Y, Miao Q S, Wei G H, Dong M M, Dong C. 2017. Quality control methods and application for the oceanic station observed data in the delayed mode. *Ocean Dev*. *Manag*., **34**(10): 109-113. (in Chinese with English abstract)

Yu T, Han G J, Guan C L, Geng Z G. 2010. Several important issues in salinity quality control of Argo float. *Mar*. *Geod*., **33**(4): 424-436, https://doi.org/10.1080/01490419.2010.518496.