Available online at www.sciencedirect.com
ScienceDirect
Ocean Engineering 34 (2007) 234–246
OCEAN ENGINEERING
www.elsevier.com/locate/oceaneng

Data quality check procedures of an operational coastal ocean monitoring network

Dong Jiing Doong[a,*], Shen Hsien Chen[b], Chia Chuen Kao[a], Beng Chun Lee[c], Sun Pei Yeh[a]

[a]*Coastal Ocean Monitoring Center, Research Center of Ocean Environment and Technology, National Cheng Kung University, 1, Ta-Hsueh Rd., Tainan, Taiwan 701, ROC*
[b]*Department of Construction Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, ROC*
[c]*Department of Environmental and Hazards-Resistant Design, Huafan University, Taipei, Taiwan, ROC*

Received 12 September 2005; accepted 19 January 2006
Available online 19 April 2006

Abstract

Field oceanographic and meteorological data are required for ocean engineering. In response to the requirement of field data, an operational coastal ocean monitoring network was established around Taiwan coast, including nine deep-water data buoys, one shallow-water pile station, 10 coastal weather stations and 10 tide stations. Data quality check procedures are necessary to ensure the accuracy of measurements. This paper presents the data quality check procedures on ocean wave data which includes automatic and manual check procedures. The checking criteria are derived using statistical theory in this paper. In addition, a sea-state-dependent algorithm is presented in this study in order to derive checking criteria of time-continuity check. It is showed to have better performance of picking up suspicious data than using fixed threshold process. This data quality check program is now used on the operational monitoring network.

Keywords: Data quality check; Oceanographic data; Data buoy; Markov process

1. Introduction

Taiwan Island locates in the subtropical region, where severe seas triggered by typhoons in summer seasons often result into terrible losses of the human life and property in the coastal areas. In order to forecast the severe sea-state for coastal hazard mitigation and make correct policy for coastal area management, the ground truth field data are required. The Coastal Ocean Monitoring Center (COMC) was therefore established under the National Cheng Kung University in 1998 to assist the government to develop and operate a hydrological monitoring network around Taiwan coast. Presently, the network consists of nine deep-water buoy stations, one shallow-water pile station, 10 coast weather stations and 10 tide stations. The location map of the stations is shown in Fig. 1. At deep-water buoy station, a 2.5 m wave-following discus buoy is deployed. The buoy is equipped with a tri-axial accelerometer to measure surface wave particle movements for the estimation of directional wave spectrum (Kao et al., 1999). At shallow-water pile stations located in areas of mild slope and sandy seabed, an ultrasonic wave gauge array is installed to provide measurements of sea surface displacements. In addition, the wind speed, wind direction, air and water temperatures, barometer pressure are also measured simultaneously. The in-situ meteorological and oceanographic observations from the network provide the government with critical information to prepare severe weather warnings. Long-term data from the network are used to calibrate and validate marine weather forecasting models and to develop design criteria for coastal structures.

Data quality control is based on both objective criteria and human experience. If incorrect or missing measurements are not properly corrected, they may significantly mislead the weather forecasting and the design conditions of constructions. The consequences of inaccurate observations may be more devastating than the lacking of observations (Gilhousen, 1988). The general data quality control includes the research and development (R & D) of

*Corresponding author. Tel.: +886 6 2744058x19; fax: +886 6 2098853.
E-mail address: doong@mail.ncku.edu.tw (D.J. Doong).

0029-8018/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.oceaneng.2006.01.011

Fig. 1. Locations of coastal ocean monitoring stations around Taiwan.

monitor technologies, daily data quality check (QC) and long-term data quality assurance (QA) as shown in Fig. 2. QC is the regulation of quality performance against set standards and acting on those whose performance is below default criteria. It must start from the time the sensors detect the objectives. QA is the activity and proof showing that the quality operation is being carried out adequately and assuring user's confidence and satisfaction in using the data. QC and QA monitor the performances of measurement systems, which is useful for scheduling maintenances and calibrations. R & D is then the improvement of new monitoring technologies. QC, QA and R & D are highly correlated and complementary R & D, daily QC and long-term QA improves the data quality. In this paper, the data QC procedures are reported.

In general, based on means of execution, the QC program is divided into automated (labeled as AutoQC) and manual (labeled as ManuQC) procedures. The AutoQC uses computer algorithms to examine a large amount of measurements. The ManuQC is then applied to the suspicious data identified by the AutoQC for further check. The algorithms of AutoQC are based on both objective criteria and subjective experiences. The use of algorithms by computers can significantly reduce the manpower that can be dedicated to the ManuQC. Based on the sequence of execution, AutoQC consists of two stages. The first stage is to examine raw time series (TSQC), which includes the default basic check (DBC) and data gap interpolation (DOC). The second stage is to examine statistical parameters (SPQC) derived from the raw time series data, which includes range rationality check (RRC), variation continuity check (VCC) and physical correlation check (PPC). This paper describes the development and application of the algorithms of the AutoQC. The AutoQC

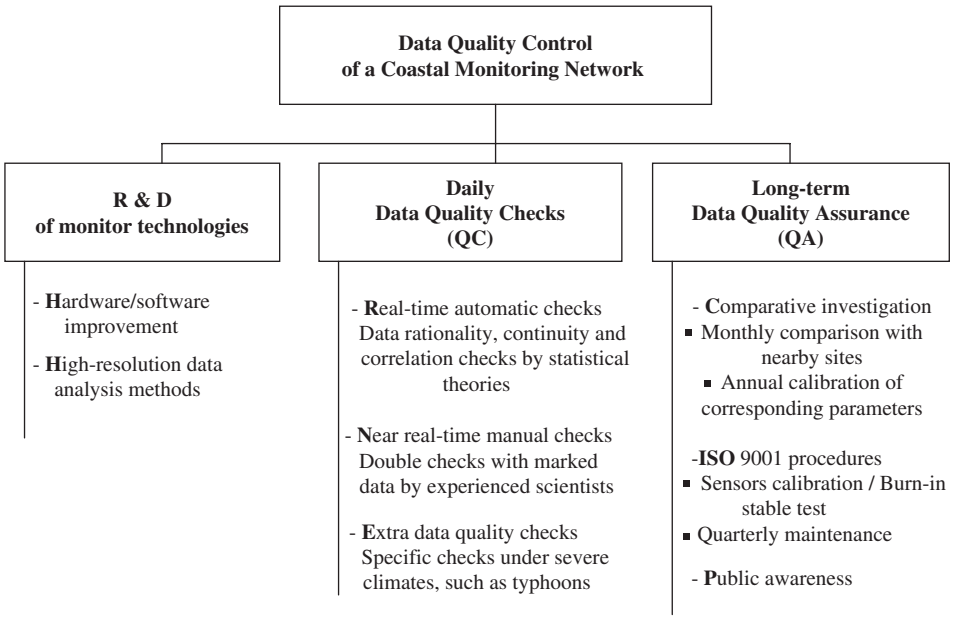


Fig. 2. Contents of the data quality control.

on the raw time series of wind and wave measurements is presented in Section 2. Section 3 describes the AutoQC on the statistical wave and wind parameters. The conclusion is given in Section 4.

2. AutoQC on time series data

Oceanographic and meteorological observations start from data acquisition in time. A poor quality time series data, such as too many erroneous values or large amount of missing data, will result into inaccurate statistical parameters. Hence, AutoQC is first applied to raw time series and then to statistical parameters derived from the time series data.

2.1. Data outlier filtering

The strategy of AutoQC is not to reject data but to locate suspicious data for further checks because they are not necessary error data. A data outlier is defined as a measurement having an erroneous value far from the rest measurements in a time series record. The data outliers are caused by various factors, which may depend on the measurement systems. For example, the data outliers can be caused by noisy ultrasonic echoes induced by the white cap from water vapor or break waves during high winds. Aging sensors can often induce signal spikes causing data outliers. Statistical parameters such as maximum wave height and significant wave height derived from the raw time series containing such erroneous measurements can be significantly overestimated. The data outliers must be identified and corrected or removed before it can be used for the calculation of statistical parameters. The data outlier can be divided into system outlier and general outlier. The system outlier is its measurement value clearly exceeding the limitations of the measurements systems or environmental conditions. These system outliers with such obviously unreasonable extreme values can be detected very easily. The general outliers are measurements within the limitations but still look suspicious due to its rather larger deviations from the rest of the measurements. The detection of such outlier often is a major challenge to the AutoQC. As an example, an ocean surface elevation time series from the Cigu pile station is shown in Fig. 3. This station is located in shallow waters with a depth of 15 m. The sea surface elevation is obtained using four ultrasonic wave gauges (at 2 Hz sampling rate) measuring distance from the sensors to sea surface. The distance between the sensor and sea surface is approximately 10 m at this pile station. Several system outliers were identified in this time series data judged by its values far exceeding the 10-m range. Also noticed in the time series, there are measurements within the 10-m range but still look suspicious, which are the general outlier marked by arrows in Fig. 3. The development of an algorithm to detect such outliers is discussed in the followings.

James (1993) presented a method to determine the outlier in a time series based on that deviations of measurements from its mean value should vary smoothly and follow a uniform distribution. The outliers can be detected when its deviations exceed the pre-determined range in the ranked deviation series. It was proposed to use three times standard deviation of time series measurements as the upper and lower limits to check the existence of data outliers. However, the multiple of standard deviation has correlation with sample sizes and confidence level for this statistical test. Besides, this approach performs best when the time series follows the normal distribution. In this

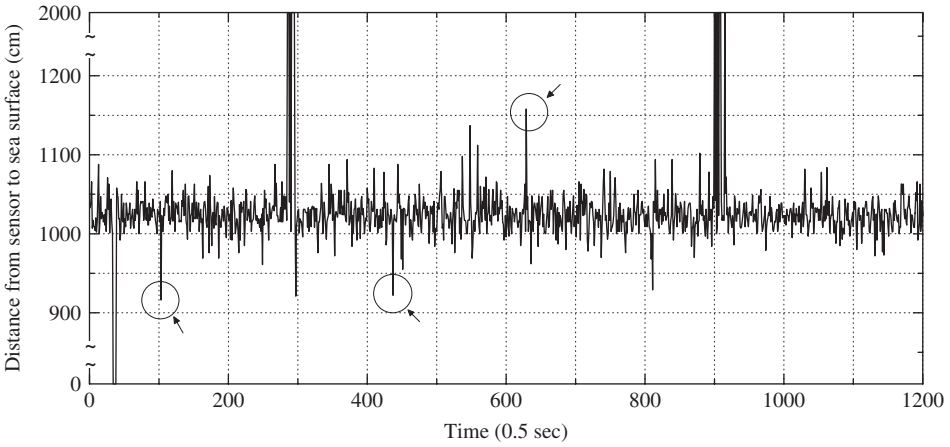


Fig. 3. Typical time series with outliers.

study, we proposed a data outlier test algorithm consisting of an upper, y_H and lower limits, y_L, which is expressed as

$$y_{\substack{H\\L}} = \hat{y} \pm K_n \cdot s_y, \tag{1}$$

where $y = \log(x)$ is the logarithmic transformation of the time series data x, which has a better fit to normal distribution; s_y is the standard deviation of y. Considering a short-term observation may be affected by a long-term decreasing or increasing trend (such as a short-term water level records affected by tidal variations), this study replaced the mean $\bar{y}$ by a long-term trend representative equation $\hat{y}$ that is calculated by linear regression equation. The coefficient K_n is related to confidence level and sample size of the measurements. For 90% confidence level, K_n can be estimated by

$$K_n = 0.49106 \cdot \log n + 1.4059, \tag{2}$$

where n is the sample size of the time series. The correlation coefficient of the regression equation is 0.9. The test was applied to check the time series in Fig. 3 with 90% confidence level. We found 4 general outliers in addition to the 21 system outliers. This result is close to James's method that found 6 general outliers and 21 system outliers as shown in Fig. 4. Since both presented algorithm and James's method have similar results, however the low and high thresholds should be defined manually for each case in James's method. It is not fit to operational requirement. The data outlier QC algorithm presented in this study uses the upper and lower limits to check the outliers, which can be programmed easily in an operational data collection network to give effective check for outliers in the raw time series.

2.2. Missing data interpolation

A continuous time series is needed for the computation of statistical parameters and spectra. To fill data gaps in the time series due to data outliers or missing measurements, a data interpolation is needed. The principle for data interpolation is to retain the statistical characteristics of the raw data as much as possible.

(*a*) *Auto-regressive (AR) Model*: The time series model provides a stable and accurate interpolation of missing hydrologic data. The autoregressive (AR) model is a simplified formula of autoregressive moving average (ARMA) model, which is commonly applied to stationary stochastic process (Box et al., 1976). The formulation of Pth-order model, AR(P) is as following.

$$\eta_t = C + \phi_1\eta_{t-1} + \phi_2\eta_{t-2} + \cdots + \phi_p\eta_{t-p}, \tag{3}$$

Where η_t is a time series, C is a white noise satisfying normal distribution and $\phi_p$is the Pth auto-regression coefficient. The model identification, parameter estimation and the following diagnostic checking are applied to establish an AR(P) model.

The type of time series model can be identified by the shapes of autocorrelation function (ACF) and partial autocorrelation function (PACF). An AR(P) process has the features of exponentially decaying on ACF figure and cuts off after order P on PACF figure. The formulas of ACF (ρ_k) and PACF (Φ_{kk}) are indicated in Eqs. (4) and (5), respectively. In the equations, μ_n is the mean of samples $\{\eta_t\}$, N is the sample size

$$\rho_k = \frac{\sum_{t=1}^{N-k}(\eta_t - \mu_\eta)(\eta_{t+k} - \mu_\eta)}{\sum_{t=1}^{N}(\eta_t - \mu_\eta)^2}, \tag{4}$$

$$\Phi_{kk} = \begin{cases} \rho_1 & k = 1, \\ \dfrac{\rho_k - \sum_{j=1}^{k-1}\Phi_{k-1,j}\cdot\rho_{k-j}}{1 - \sum_{j=1}^{k-1}\Phi_{k-1,j}\cdot\rho_j}, & k > 1. \end{cases} \tag{5}$$

Assume the order of model is N. Eq. (3) can be expressed in a matrix formation as Eq. (6). The parameter matrix [Φ] in Eq. (6) can then be solved by applying the least

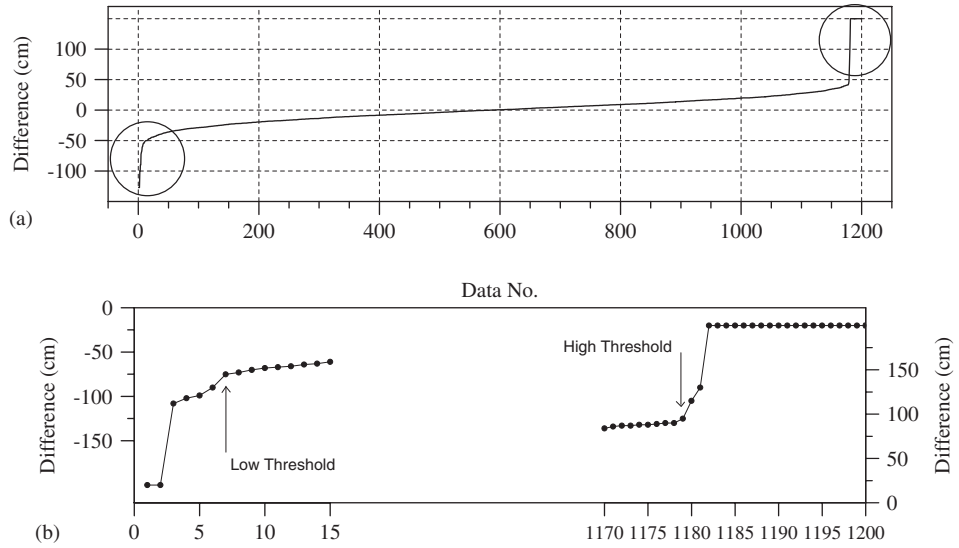


Fig. 4. (a) Distribution of difference of each record in the time series to its mean by James's method. (b) The detail plots at the margin of upper figure.

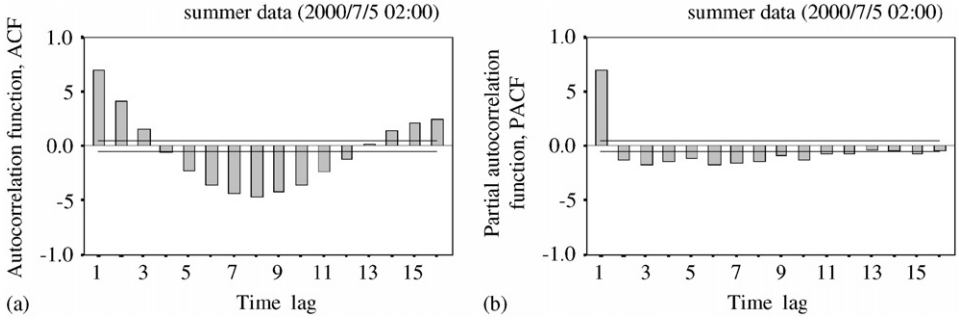


Fig. 5. (a) ACF of one summer data, (b) PACF of one summer data.

square method

$$\begin{bmatrix} \eta_{p+1} \\ \eta_{p+2} \\ \vdots \\ \eta_N \end{bmatrix} = \begin{bmatrix} 1 & \eta_p & \eta_{p-1} & \cdots & \eta_1 \\ 1 & \eta_{p+1} & \eta_p & \cdots & \eta_2 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & \eta_{N-1} & \eta_{N-2} & \cdots & \eta_{N-p} \end{bmatrix} \begin{bmatrix} C \\ \phi_1 \\ \vdots \\ \phi_p \end{bmatrix}. \quad (6)$$

The time series model should be diagnosed by parameter test and lack-of fit test. The T-ratio value of parameter test is to examine if the parameters are significant enough to prevent overfitting. In order to assess the degree of fit between observed $\{\eta_t\}$ and calculated values by model $\{\hat{\eta}_t\}$, a parameter named as non-dimensional root mean square error (NRMSE) was defined as

$$\text{NRMSE} = \sqrt{\frac{1}{N-1}\sum_{t=1}^{N}\left(\frac{\eta_t - \hat{\eta}_t}{\eta_t}\right)^2}. \quad (7)$$

The wave climate around Taiwan coastal waters is seasonal varied significantly. In this study, 10 time series data sets of surface waves from summer and winter seasons respectively, was used, and another 10 time series data set from severe seas caused by typhoons was used for calibration of AR model. The wave time series data are from the Cigu pile station. The selected data sets have no outlier and tidal effect. Figs. 5(a),(b) and 6(a),(b) show an example of ACF and PACF plots from summer and winter, respectively. AR(1) model was identified for these two examples based on the fact that the exponentially decaying on ACF plot and cut-off PACF plot after first time lag. On the other hand, we found the model identification for typhoon data varies depending on individual case. Although, ACF and PACF plots from one time series during typhoon Bilis in August 2000 are similar to the result of summer data as well as AR(1) model (see Fig. 7(a) and (b)). However, the analyzed time series during typhoon Yagi is difficult to be identified to AR model (see Fig. 8(a) and (b)). Identification and diagnostic checks of all summer and winter data sets are listed in Table 1. The wave records from summer and winter belong to AR(1) model with most of the T-ratio values of calibration cases greater than 2.0 indicating significant

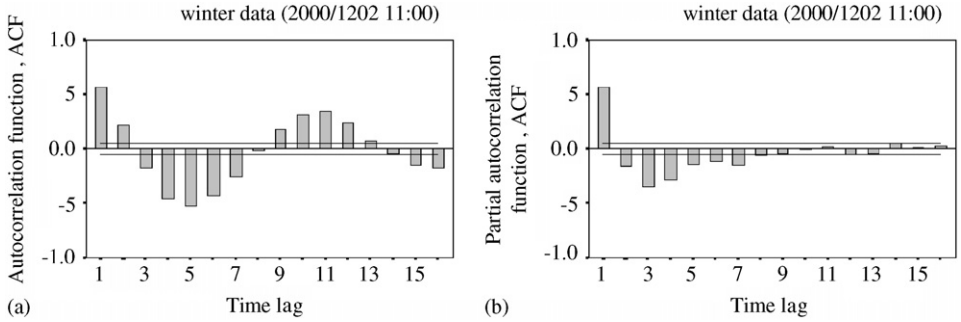


Fig. 6. (a) ACF of one winter data, (b) PACF of one winter data.

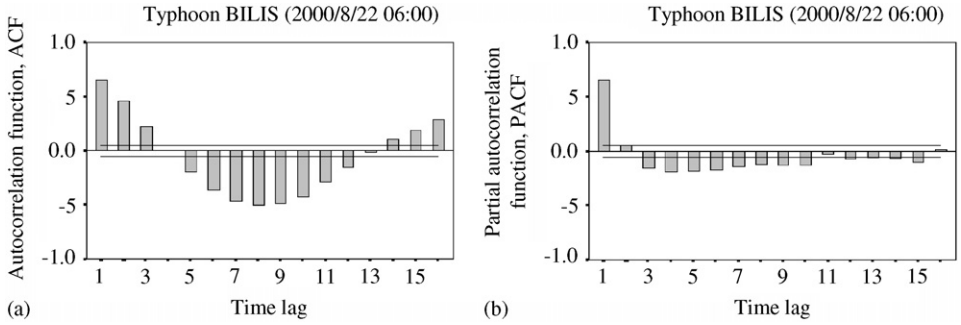


Fig. 7. (a) ACF of one typhoon BILIS data, (b) PACF of one typhoon BILIS data.

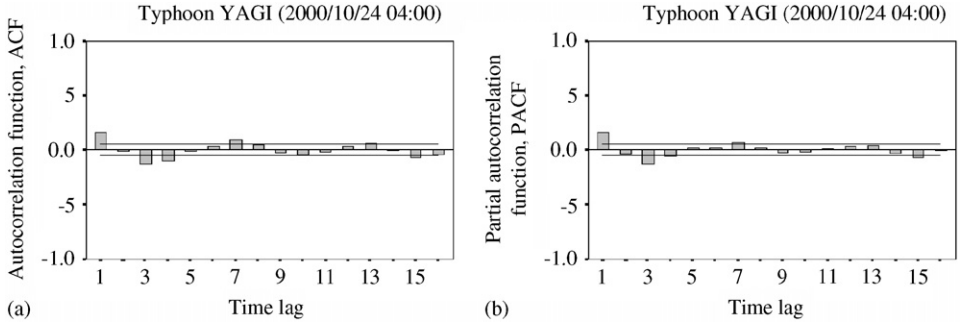


Fig. 8. (a) ACF of Typhoon YAGI data, (b) PACF of Typhoon YAGI data.

Table 1
Identification, parameter estimation and diagnostic check of AR model

Data group	Model identification		Parameter estimation		Diagnostic check	
	Calibration cases	Group assessment	ϕ_1 (mean/sd./COV)	C	*T*-ratio	NRMSE
Summer data	9@AR(1)1@ none	AR(1)	0.667/0.033/4.9%	340.8/41.6/12.2%	31.1	2.5%
Winter data	8@AR(1)2@ none	AR(1)	0.486/0.078/16.0%	527.6/69.5/13.2%	19.5	3.1%

Note: sd. = standard deviation, COV = coefficient of variation.

parameter. The average NRMSE is less than 5% showing good agreements of AR(1) model on the data. It is therefore concluded that the wave records in summer and winter seasons can be simulated by AR(1) model at Cigu station. The coefficient of variance (COV) for the model parameters are below 20%, means the spreads of model parameters are accepted to use the mean values to be the regional model parameters for summer and winter data at

Cigu station. The regional models are then applied to fit with other hundred data sets from the same Cigu station given the average NRMSE of 12.8%. It presents the acceptable results of using AR(1) model to simulate time series data in summer and winter.

(b) Limitation on data interpolation: To assure data gap can be reasonably interpolated without the loss of its connatural characteristics, the amount of missing data in a time series record should be limited. The maximum amount of missing data allowed in a record depends on the data patching ability of the chosen interpolation methods to return its original statistical characteristics. The absolute bias of statistic (ABS) parameter is used to judge the bias of statistical parameters after interpolation, which is expressed as

$$\mathrm{ABS} = \left| \frac{S - \hat{S}}{S} \right| \tag{8}$$

where S is the statistical parameter from original data series (such as significant wave height, period), and $\hat{S}$ is the same statistical parameter from the data series interpolated by the AR(1) model for a specified amount of data lost. To determine the maximum amount of missing data of a time series, Monte Carlo method was applied to simulate missing data of typical summer and winter surface waves time series. The random positions were generated from a uniform distribution for 100 runs. Totally 1200 samples are simulated in each run. Fig. 9 show the ABS of significant wave height versus the amount of lost data in the simulated sea surface elevation time series. The simulations show that ABS increase as the total amounts of missing data increase. Assessment by the ABS of other wave parameters from the simulated time series showed similar results. It is empirically determined that 20% bias (ABS = 0.2) should be maximum allowable level of any wave parameters, which means the absent data amount should not exceed $\frac{1}{6}$ of the total observation data.

3. AutoQC on statistical parameters (SPQC)

The quality check of statistical parameters use algorithms based on limitations imposed by measurement ranges, temporal variability, and correlations among parameters.

3.1. Range-rationality check (RRC)

Any statistical parameters cannot have its value exceed the range of sensors or the physical restrains of the marine environments. The range-rationality check is to assure that the magnitudes of statistical parameters are first within the limits. For example, due to the wave breaking induced in shallow waters, the measured significant wave height cannot exceed the breaking wave height imposed by the local water depth. In addition, the significant wave height should not be over the measurement range of equipments.

3.2. Variation-continuity check (VCC)

The variation-continuity check consists of time-continuity check (TCC) and space-continuity check (SCC), which are based on the concept that evolution of natural phenomenon in time or space should be gradual and smooth. The National Data Buoy Center (NDBC) of United States has developed three TCC algorithms for pressure, temperature, and other parameters (NDBC, 2003). In this study, it is focused on the TCC algorithm

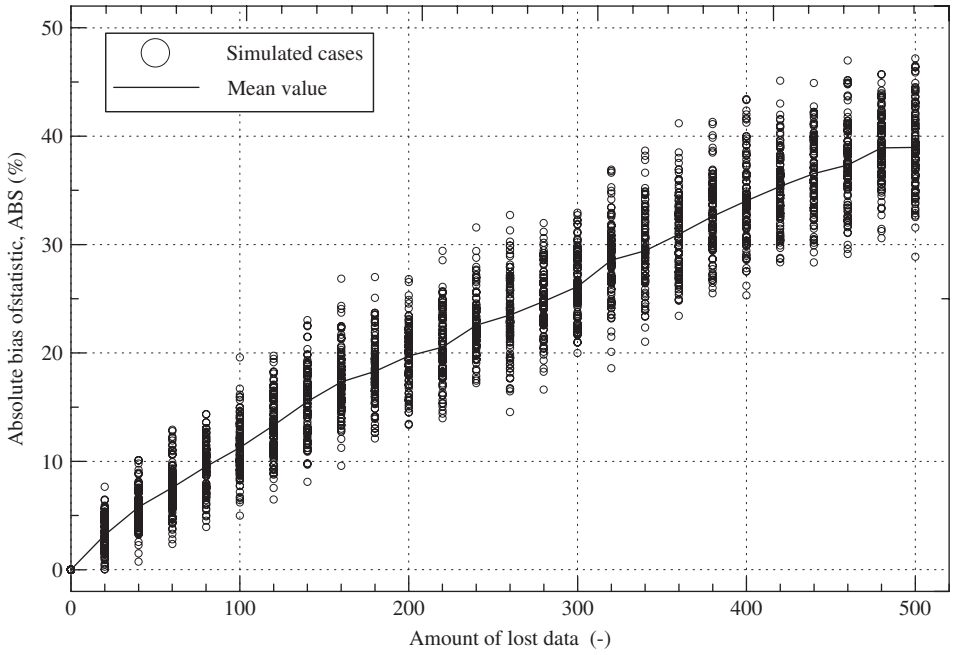


Fig. 9. Bias of significant wave height from interpolated time series versus amount of lost data.

on significant wave height. Field data show that the upcoming sea state has a leaving effect from sea states of preceding hours. This implies that the temporal change of significant wave height depend on the significant wave heights of previous hours. A sea-state independent fixed threshold for the TCC of significant wave height could underestimate the temporal variability in high seas and overestimate the variability in low seas. This paper presents a sea-state-dependent TCC algorithm for significant wave height which is developed based on the property of Markov process in statistics.

(a) Markov process of significant wave height: Stochastic time processes can be ranked in increasing order of complexity, depending upon the degree of causality they embody as having a sort of "memory" of its own past. That is, the random event which occurs at time n may be dependent upon that which occurred at time $(n-1)$ or earlier stage. Markov process is a process with a short-term memory that means each random event is only influenced to some degree by its previous predecessors (Ang and Tang, 1975). Markov process has no direct memory of earlier events. This study examines processes with simpler first order model for the operational QC program. That is, it may be possible to predict the probability of states of significant wave height at time n refer to its formal sea state at time $(n-1)$. The acquiring additional information on time $(n-2)$, time $(n-3)$, etc., may not provide further useful information for making predictions at time n.

If a hydrologic data x at time n is affected by its previous state x_{n-1}, we can model the stochastic process by the conditional probability $P[x_n|x_{n-1}]$. Furthermore, the transition probability of the data change from state i in stage X to state j in stage Y is expressed in the following equation:

$$p(j,i) = P[Y = y_j|X = x_i] \tag{9}$$

That is, the probability that the information in stage Y (i.e. time $n+1$), given the knowledge that it was in stage X (i.e. time n). It is assumed that the transition mechanism of the system, although random, remains constant over time, i.e. called the homogeneous Markov process. This collection of probabilities forms a transition probability matrix. Divide the history data into i and j non-overlapping states, respectively, in the sequent stages, the transition probability matrix can then be expressed in the following equation.

$$P_{Y,X} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1j} \\ p_{21} & p_{22} & \cdots & p_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ p_{i1} & p_{i2} & \cdots & p_{i,j} \end{bmatrix} \tag{10}$$

$$p_{i,j} = P\Big[Y_{n+1} = y_j|X_n = x_i\Big] = m_{i,j} \Big/ \sum_{k=1}^{j} m_{i,k} \tag{11}$$

In the above equation, $m_{i,j}$ is the sample number happened from state i of stage X to state j of stage Y.

(b) Built and rebuilt of transition probability matrix: As an example for transition probability matrix analysis, significant wave height measurements at 2-h interval from Longdong buoy in the year 2000 are selected. There are 4392 significant wave height data, which are then divided into 10 states (0–30, 30–50, 50–80, 80–100, 100–150, 150–200, 200–300, 300–400, 400–500, >500 cm). The resulted transition probability matrix is shown in Table 2. From the table, we found most of the high probability events occurred at the same states between two sequent time steps indicating the fact of large waves always comes after large waves and vice versa. In order to fit this method to data QC process, the second stage of Markov process was modified in this study. As the observation interval of significant wave height is 2 h, the significant wave height variation in 2 h is defined as the variable of second stage. This newly calculated sea state variations are divided into 10 intervals, which are 0–10, 10–20, 20–30, 30–40, 40–50, 50–60, 60–70, 70–80, 80–90 and >90 cm. The rebuilt transition probability matrix was calculated and shown in Fig. 10. The critical limitations of the following stage were

Table 2
Transition probability of significant wave height between different states

Next states Present states	State I	State II	State III	State IV	State V	State VI	State VII	State VIII	State IX	State X
	0–30	30–50	50–80	80–100	100–150	150–200	200–300	300–400	400–500	>500
State I 0–30	46.9	51.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
State II 30–50	7.6	76.5	15.6	0.3	0.0	0.0	0.0	0.0	0.0	0.0
State III 50–80	0.1	12.3	67.4	15.8	4.3	0.0	0.0	0.0	0.0	0.0
State IV 80–100	0.0	0.2	30.3	43.9	24.2	1.1	0.2	0.0	0.0	0.0
State V 100–150	0.0	0.0	2.3	16.2	63.7	16.1	1.7	0.0	0.0	0.0
State VI 150–200	0.0	0.0	0.0	0.4	26.2	52.1	21.0	0.4	0.0	0.0
State VII 200–300	0.0	0.0	0.0	0.0	1.5	23.5	68.0	6.3	0.7	0.0
State VIII 300–400	0.0	0.0	0.0	0.0	0.0	1.2	42.7	51.2	4.9	0.0
State IX 400–500	0.0	0.0	0.0	0.0	0.0	0.0	9.5	28.6	52.4	9.5
State X >500	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	66.7	33.3

Unit of wave height: cm.

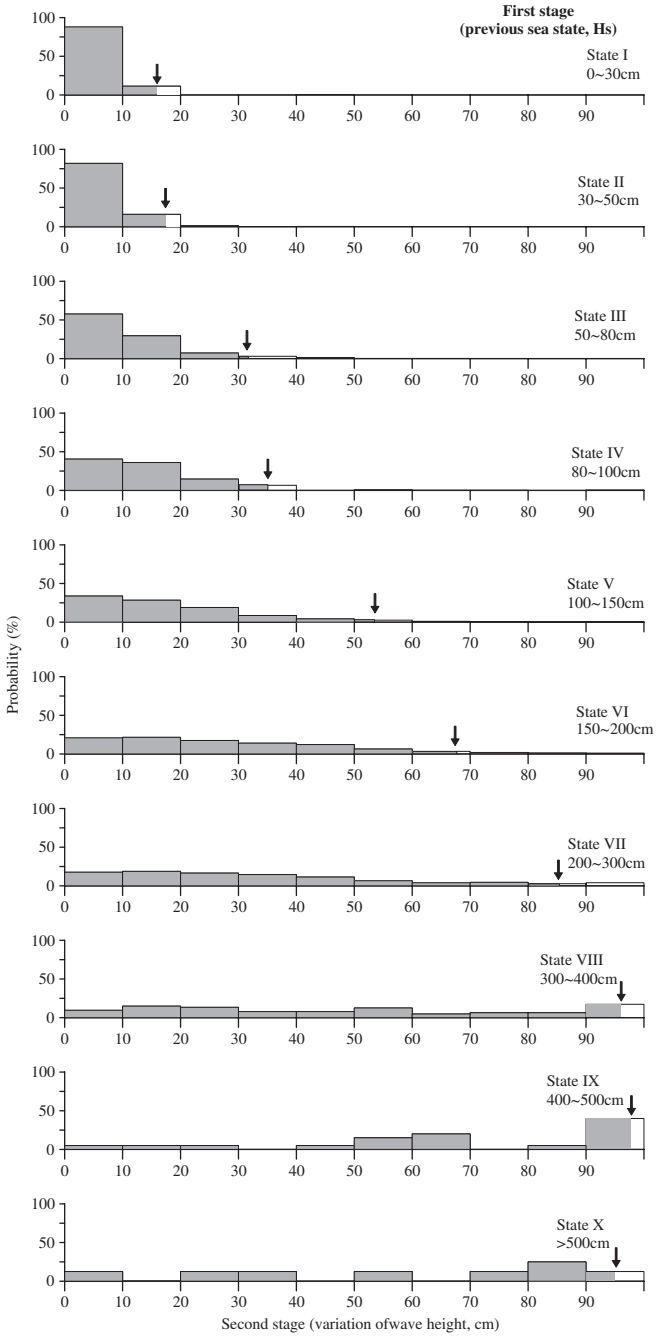


Fig. 10. Rebuilt transition probability matrix of significant wave height.

interpolated under a conference level of 95% (shadow area in Fig. 10). These values will be used as the allowable range of sea state change in 2 h for TCC. The values are listed in Table 3 according to the previous sea state. For example, if the current significant wave height is 100–150 cm, there are 95% of probability that the change of significant wave height in the following hour is within 53.0 cm.

Comparison of the checking performance of presented sea-state-dependent method of TCC process with fixed threshold was studied. The fixed thresholds of TCC process used by NDBC is calculated by the formula

Table 3
Allowable variation of wave height of different sea states

Previous sea state (wave height, cm)	Allowable variation (2 h lag time)
0–30	16.1
30–50	18.2
50–80	31.3
80–100	35.2
100–150	53.0
150–200	68.5
200–300	85.2
300–400	169.5
400–500	188.9
>500	157.8

Table 4
Marked numbers of different failed QCs and measurements

Types of QC	Measurements			
	Wave height	Wind speed	Air temperature	Air pressure
Out-TSQC	999.8	99.8	99.8	9999.8
Out-RRC	999.1	99.1	99.1	9999.1
Out-VCC	999.2	99.2	99.2	9999.2
Out-PCC	999.3	99.3	99.3	9999.3
Out-ManuQC	999.9	99.9	99.9	9999.9

$\sigma_T = 0.58\delta\sqrt{T}$, where σ_T is the allowed difference after T hour. The parameter, δ is a non-dimensional parameter, its value varies with the changed in observation objects. This value varies upon the local wave climate and 6.0 is used for significant wave height by NDBC. By applying the above equation and parameter, the TCC threshold of wave height for 2 h is up to 4.92 m which is not a proper value to use for the wave climate around Taiwan. So, the threshold was reduced to 0.492 m in order to have a reasonable comparison. The given adjustable criteria of presented Markov method is listed in Table 4. Comparison result of TCC by TCC criteria and NDBC formula for significant wave height in the year 2000 is shown in Fig. 11. Presented method filters out 156 records of wave height that does not comply with the continuity principles while NDBC filters out 344 suspicious data. The amount of suspicious data is overestimated by NDBC, which is twice of presented algorithm. This difference could be even larger in severe sea states during winter monsoon seasons around Taiwan. Because the suspicious data by the AutoQC should be re-checked by ManuQC, less suspicious data could reduce the job of followed ManuQC.

3.3. Physical-correlation Check (PCC)

Oceanographic and meteorological parameters provide measures of various physical elements of marine environment,

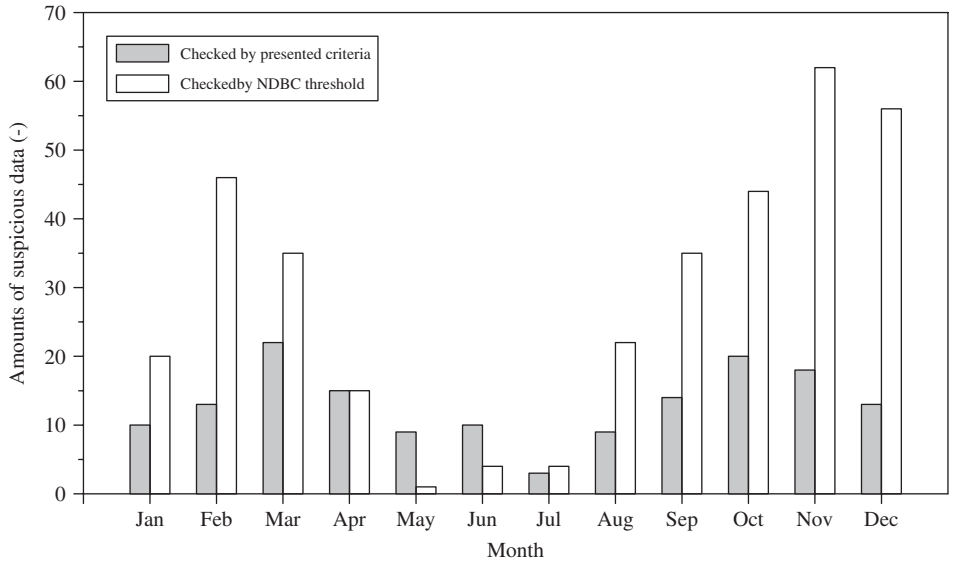


Fig. 11. Comparative result of time-continuity check by present sea-state-dependent algorithm and NDBC fixed threshold method.

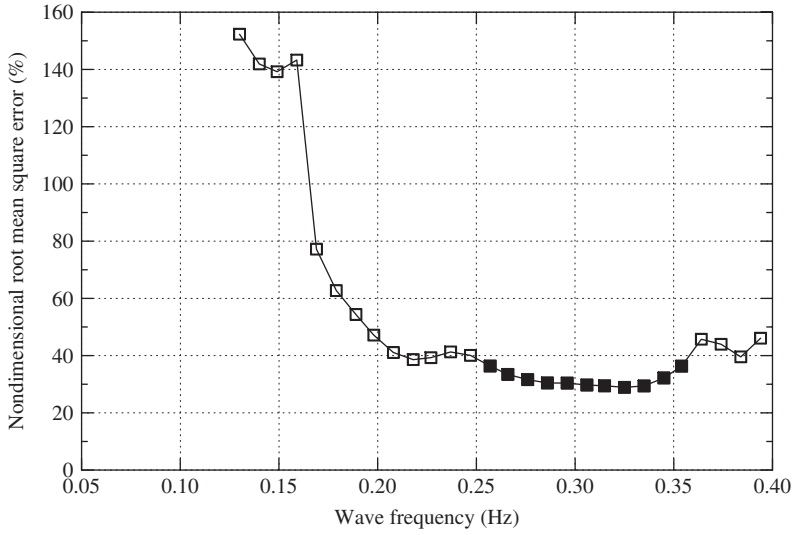


Fig. 12. Correlation between mean wind speed and wave energy within different frequency bands.

which are often closely related. The correlation among various parameters can be used to develop algorithms for QC (i.e., physical correlation check (PCC)). For example, sea surface winds are the major generation source of wind waves, a close relationship between wave energy and wind speed can be expected. Steele and Marks (1979) show that local wind is strongly correlated with wave energy in the frequency of 0.2–0.27 Hz. Lang (1987) showed a better wind-wave correlation using the square of the mean wind speed 4 h prior to the observation.

In this study, PCC algorithm on wind and wave data based on the correlation of wind speed and high-frequency wave energy is presented. This study used 3725 simultaneous wind and wave data sets from Longdong buoy in the year 2000. In order to study wind-wave correlation, simultaneous steady wind and wave data are selected based on following three criteria (1) mean wind speed less than 25% of variance within a continuous 8 h period, (2) the difference between wind and wave directions are within 90°, and (3) the wave spectral peak frequency is higher than the peak frequency based on the PM spectral model for the given wind speed.

Due to the scatter of the data, this study used again the NRMSE between data and its representative from regression equation instead of the correlation coefficient to describe the performance of the regression analysis between wind and wave. As showed in Fig. 12, good relations of wind and wave exist in the frequency bands of 0.257–0.355 Hz. According to the corresponding mean wind speeds, wave energy over frequencies bands 0.257–0.355 Hz are then divided into 14 groups regions from 0.2 m/s to more than 25 m/s. Regression analysis

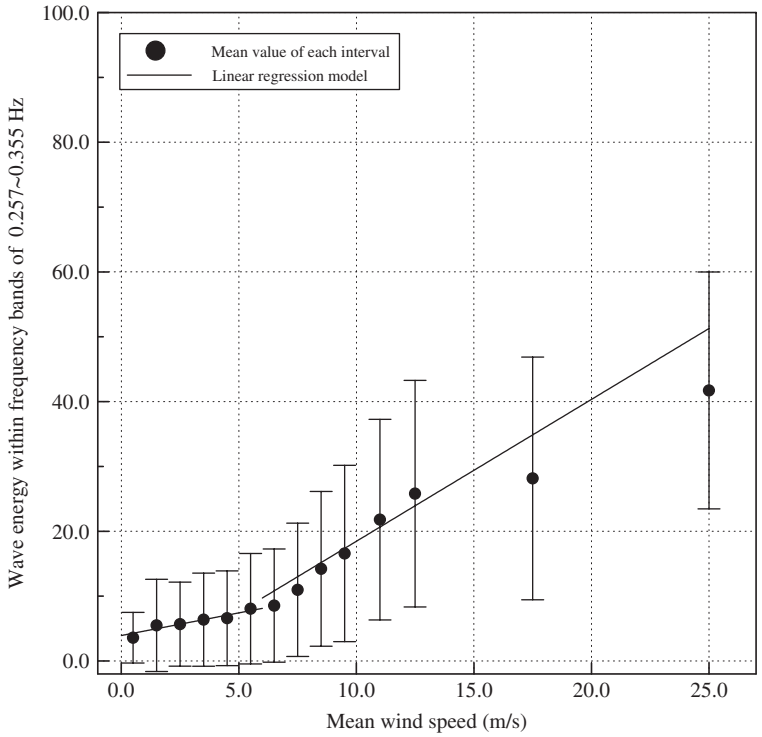


Fig. 13. Correlation between mean wind speed and wave energy within the frequency band of 0.257–0.355 Hz.

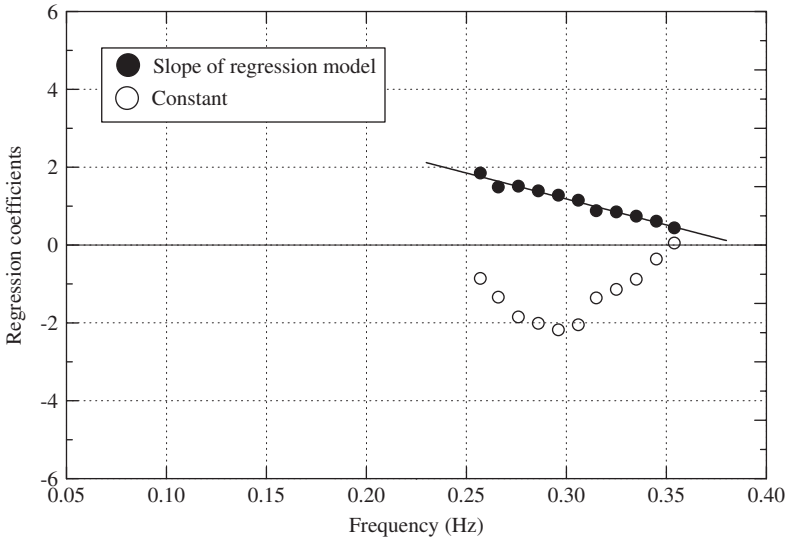


Fig. 14. Regression coefficients of wind-wave correlation versus wave frequencies.

results indicated linear relation exists between wind and wave energy when wind is greater than 6 m/s, specially in the range of 6–15 m/s. There is no significant relation exists between wind and mean wave energy when the mean wind speed is less than 6 m/s. Fig. 13 show the relation between mean wind speed and wave energy in the frequency range of 0.257–0.355 Hz, this relation can be expressed as

$$E(\Delta f) = b(f) \cdot U_{10} + a, \tag{12}$$

where a and b are the regression coefficients. The PCC on wave data is to assure wave energy at frequency bands of 0.257–0.355 Hz is within $\pm 10\%$ of that estimated by Eq. (12). The slope of the linear relationship between wind and wave energy at each frequency band is shown in Fig. 14. This slope represents the transmitting rate of wind energy to the wave. Fig. 14 shows the slope decreases with an increasing frequency, which implies low-frequency waves are less acceptable to wind energy input than

high-frequency waves. This analysis was obtained from the data at a 2.5-m discus buoy, indicating the response of wave from wind field by this type of buoy.

The significant wave statistical parameters are calculated from wave spectra, such as significant wave height $H_s = 4.004\sqrt{m_0}$, where m_0 is total wave energy. Therefore, when the significant wave height is between 3.6 and 4.4 times of root total energy under a 90% confidence level, the data is viewed as a valid data. Otherwise, the wave height data will be marked by the PCC system.

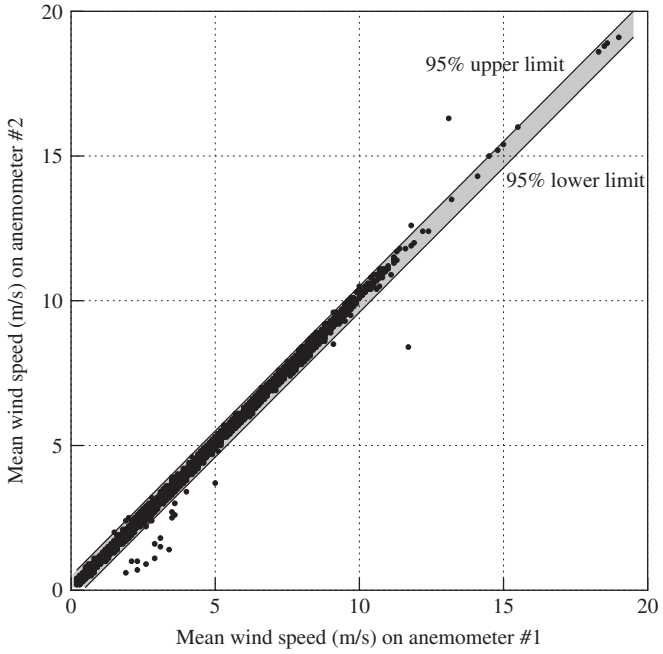
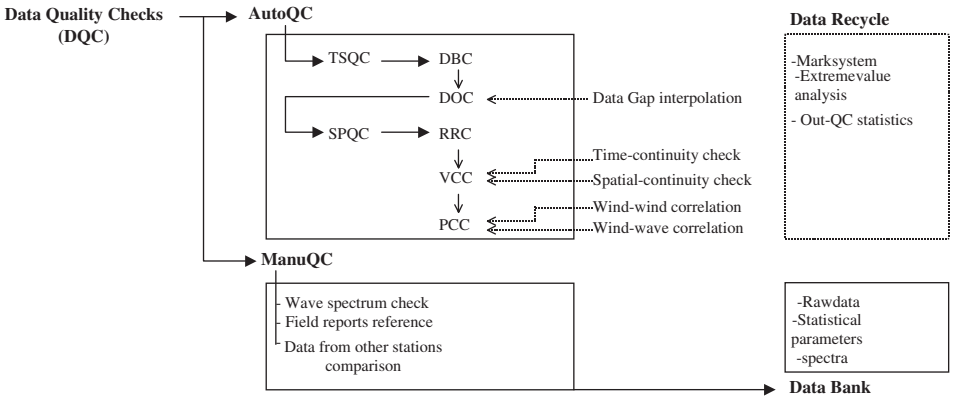


Fig. 15. Cross-correlation between two anemometers at Longdong buoy.

In addition to the correlation between the wind and wave, we often can compare measurements from two collocated sensors measuring the same parameters. For example, two anemometers are often installed on the data buoy or pile station to assure the acquisition of wind data and reduce the probability of equipment malfunction. Quality of wind measurements of the two anemometers can be checked by the comparison between them, which can also be used to show the deviation caused by aging or damaged anemometers. This study analyzed the relationship between the average wind speeds in the year 2000 from the two anemometers installed on Longdong buoy. The result of regression analysis is shown in Fig. 15. The averaged wind speeds from the two anemometers appeared to be relatively synchronized, showing the stability of the observation system and increasing the reliability of the wind speed data. The upper and lower limits of the 95% confidence interval of the linear regression equation are the check thresholds of PCC on the wind speed. Wind speeds which lie outside the band region will be treated as data failing QC.

4. Conclusion

Meteorological and oceanographic observations from a network of moored buoys and fixed platforms in the coastal waters of Taiwan are used to validate and improve marine weather forecasting models and to be the design criteria of coastal and ocean engineering constructions. Erroneous or missing measurements caused by severe seas, human errors and aging instruments could significantly decrease the values of measurements. To assure the quality of measurements, a quality check program to provide a systematical and timely examination on the measurements of the network is installed. In this paper, the developments and applications of computer algorithms used in the two-stage AutoQC



Symbols:
1) *AutoQC*: Automatic QC; *ManuQC*: Manual QC
2) *TSQC*: Time Series QC
DBC: Default Basic Check;*DOC*: Data Outlier Check
3) *SPQC*: Statistical Parameters QC
RRC: Range Rationality Check,*VCC*: Variation ContinuityCheck,*PCC*: Physical Correlation Check

Fig. 16. Flowchart of data QC procedures. Symbols: AutoQC: automatic QC; ManuQC: manual QC; TSQC: time series QC; DBC: default basic check; DOC: data outlier check; SPQC: statistical parameters QC; RRC: range rationality check; VCC: variation continuity check; PCC: physical correlation check.

procedures on wave measurements is described. At the first stage, quality check on raw time series data is performed to detect erroneous measurements based on upper and lower limits related to sample size and confidence level. The first order auto-regression model, AR(1), is used to interpolate data gaps in the time series. At the second stage, statistical parameters of wave and wind measurements are examined based on measurement ranges, continuity of temporal variations and correlations among wind and wave measurements. A sea-state-dependent variation threshold is developed from Markov process for the time-continuity check of significant wave height. The close correlation between the mean wind speed and the wave energy between 0.257 and 0.355 Hz is shown in the study and used for the quality check of both wind and wave data.

The quality control of a data network requires the use of the power from both computer algorithms and human experiences. The AutoQC is not to simply reject measurements; instead it is to identify the suspicious measurements from large amount of observations by the network for further manual check. The present data quality check procedure is now used on the operational monitoring network in Taiwan. The flowchart is shown in Fig. 16. The successful automated quality control program is to balance the need of preserving both quality and quantity of observations.

Acknowledgments

The study was supported by National Science Council (NSC 93-2611-E006-009) and Water Resources Agency of Taiwan, ROC. The in-situ data used in this paper was provided by Central Weather Bureau. The authors would like to display their great thanks.

References

Ang, A.H-S., Tang, W.H., 1975. Probability Concepts in Engineering Planning and Design, Basic Principle. Wiley, New york.

Box, G., Jenkins, G.M., Reinsel, G., 1976. Time Series Analysis: Forecasting and Control. Prentice-Hall, Emglewood Cliffs, NJ.

Gilhousen, D.B., 1988. Quality control of meteorological data from automated marine stations. Proceedingss of the fourth International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology, Miami, USA, pp. 113–117.

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

Kao, C.C., Chuang, L.Z.H., Lin, Y.P., Lee, B.C., 1999. An introduction to the operational data buoy system in Taiwan. Proceedings of the International Conference on the Mediterranean Coastal Environment, Antalya, Turkey, pp. 33–39.

Lang, N.C., 1987. An algorithm for the quality checking of wind speeds measured at sea against measured wave spectral energy. IEEE Journal of Oceanic Engineering 12 (4), 560–567.

National Data Buoy Center (NDBC), 2003. Handbook of automated data quality control checks and procedures of the National Data Buoy Center, NDBC Technical Document 03-02. National Oceanic and Atmospheric Administration, Mississippi, USA.

Steel, K.E., Marks, G.E., 1979. Detection of NDBC wave measurement systems malfunctions. Proceedings of the IEEE OCEANS '79, San Diego, CA, USA, pp. 226–236.