Quality Control of ARGO Data Based on Climatological T-S Models

Ji Fengying (纪风颖)\(^1,2\) Lin Shaohua (林绍花)\(^2\)

1 College of Environment, Ocean University of China, Qingdao 266003, Shandong, China
2 China National Marine Data & Information Service, Tianjin 300171, China

Abstract By implementing the ARGO program, a large number of T-S profiles can be observed in the world ocean. However, it is very difficult to examine changes of the sensitivity of the sensors equipped at the ARGO floats, because it is difficult to understand their condition in the sea and the reliability of the data. Quality control must be done in order to avoid the wrong conclusion deduced from the wrong data. One of the realistic methods for quality control of the ARGO data is the comparison with the local climatology. High quality climatological T-S models in northwest Pacific have been built based on the Nansen bottle data and CTD data for the quality control in NMDIS. The models are used to check the ARGO data in this area and have got good result.

Keywords ARGO program, Quality control, CTD, Nansen

Introduction

Autonomous CTD profiling floats are free-moving floats that report vertical profiles of salinity, temperature and pressure at regular time intervals. The ARGO program plans to deploy 3000 such floats to observe the upper 2000 m of the global ocean. It enables high-resolution and real-time observations of the surface and middle layers of the oceans (e.g., Mizuno, 2000). The ARGO program is part of the Global Ocean Observing System (GOOS) and will contribute to the Climate Variability and Predictability Study (CLIVAR) and the Global Ocean Data Assimilation Experiment (GODAE).

It is difficult to directly monitor the conditions of the onboard sensors of the ARGO floats once they are deployed in the ocean. Also, it is extremely difficult to determine the cause of sensor deterioration during the drift of the floats in the ocean. Therefore, to evaluate the quality of the data obtained by the ARGO floats (ARGO data) and, if possible, to obtain high-quality data are important for the program.

Therefore, we can say that the success of the ARGO project depends on the efficiency of the data quality control. Generally, the temperature and salinity in deep layers are considered to be relatively stable\(^1,2\). Therefore, the most practical method of making quality evaluations of ARGO data is to compare them for the deep layers to the local climatology. Such a method has been adopted by some organizations. For example, in the University of Washington, Wong et al. (2001)\(^3\) is developing a correction method for salinity data using profile data from the World Ocean Database 1998. In the Commonwealth Scientific and Industrial Research Organization (CSIRO) of Australia, salinity data are corrected by using the WOCE-CTD data obtained in the vicinity of floats (CSIRO, 2001).

As a country abut against NW Pacific, China is more interested in the data of this area. Besides the real time quality control method recommended by the international ARGO program, we used NODC
salinity-density models for reference, established temperature-salinity models to check ARGO data automatically in northwest Pacific, so that quality control can be carried out on a large amount of ARGO data within a limited period of time.

1 Establishing high quality climatologic T-S models

1.1 The limitation of T-S models by the bivariate frequency distributions method

Our idea about the T-S models comes from the NODC salinity-density models. NODC generated the models for each five-degree square of ocean for which sufficient data were available (10 or more valid stations). Using all valid (non-questionable) data greater than or equal to 100 m in depth, NODC compiled bivariate frequency distributions of salinity versus density (sigma-t) from those derived from the midline and envelope (upper and lower bound) models that define expected values and ranges of salinity as a function of density. Water mass model comparison checks each new station against salinity-density models derived from the historical data in the NODC oceanographic station data file (Fig. 1) [4].

Because density is derived from temperature, salinity and pressure, wrong salinity can result in wrong density. In this occasion, the salinity-density models will not be effective. Since the two main state variables of the ocean, temperature, $T$, and salinity, $S$, are related to each other by definite patterns that represent the mean characteristics of a region [5], the idea of T-S models comes out.

At first, bivariate frequency distributions method was used to set up the T-S models. In a curtain $2^\circ \times 2^\circ$ square (130° - 132° E, 20° - 22° N), temperature was taken as an independent variable, and salinity as an attributive variable and the temperature is divided into a number of sections. In each temperature section, by using the bivariate frequency distributions method we calculate the midline and the envelope value of salinity [5]. In this method, $\triangle T = 0.2$ °C, $\triangle S = 0.01$, in a certain $2^\circ \times 2^\circ$ square (132° - 134° E, 20° - 22° N), the result is shown in Fig. 2.

Because the number of observation data in each temperature section is different, the envelope of salinity changes acutely (Fig. 3). For example, in the 2.65 °C - 2.85 °C section, the number of observation data is 47, and the salinity lower bound, middle line and upper bound are 34.51, 34.57, 34.65 respectively, while in the next section, 2.85 °C - 3.05 °C, the number of observation data is 29, and the salinity lower bound, middle line and upper bound are 34.20, 34.57, 34.65 respectively. The salinity...
lower bound changes from 34.51 to 34.20 acutely. Although $\Delta T$ changes, the analogous result appears.

**Fig. 2 130° - 132° E, 20° - 22° N, T-S model by bivariate frequency distributions method**

**Fig. 3 130° - 132° E, 20° - 22° N, the number of observation data in each temperature section**

### 1.2 Establishing T-S models by multinomial curve fit

A multinomial curve fit was used to define expected values and ranges of salinity versus temperature. The detailed information is as follows.

#### 1.2.1 Historical data amalgamation

Most Chinese ARGO floats have been deployed in Northwest Pacific (0° N, 130° E - 15° N,

**Fig. 4 Northwest Pacific historical temperature and salinity data distribution map**
125° E - 23° N, 125° E - 30° N, 145° E - 0° N, 145° E). In this area, a large number of historical temperature and salinity profiles, such as CTD data, Nansen bottle data (Fig. 4) have been collected, processed and quality controlled by NMDIS, so that climatological T-S relationship background can be established by these data.

The northwest Pacific (100° - 180° E, 0° - 60° N) is divided into numerous 2° × 2° subregions by the American navy's square code.

1.2.2 Quality control of historical data

Although CTD data and Nansen bottle data have been processed through general quality control in NMDIS, this idea emphasizes the dataset climatological character, and the singular data are eliminated as follows.

1.2.2.1 Pick out singular data by Pa T rule

In each 2° × 2° square, for each observed datum \( x_i \), if \( |v_i| > 3\sigma \), the datum is regarded as singular, in which \( \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \) and \( N \) is the number of the observation value, and \( S \) is substituted for \( \sigma \), in which \( S = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} v_i^2} \).

1.2.2.2 Test singular data by the statistical F method

Because the standard error and deviation in the Pa T method are influenced by the number of the samples, the singular datum must be tested by the statistical method. Statistical F method is used to check whether the singular datum is remarkably exceptional or not.

\[
F = \frac{n + 1}{n - 1} \left( \bar{x} - x_{n+1} \right)^2 \left( \frac{n}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \right)
\]

In which, \( x_{n+1} \) is the datum whose exceptional character will be checked, and \( n \) is the number of the rest data.

\( F \) approximately obeys \( F \) distribution which freedom is \( (1, n) \). Given the upside boundary data \( \alpha = 0.05 \), look it up in the \( F \) distribution table to get the critical value \( F_{\alpha}(1, n) \). When \( n \) is greater than 50, the critical value is 4. If \( F \) is greater than 4, the datum \( x_{n+1} \) will be eliminated, or this datum will be kept in the dataset.

1.2.3 Establishing T-S models by multinomial curve fit

First, in each 2° × 2° square, all historical temperature and salinity data are mapped to the T-S coordinate system, in which temperature \( T \) is taken as the independent variable, and salinity \( S \) as the attributive variable. A polynomial curve is applied:

\[
S = f(T)
\]

which is based on the least square method to fit all the data points. From this curve, the salinity midline was got.

The points \( (T_i, S_i) \) above the line are selected to calculate the difference between \( S_i \) and \( f(T_i) \):

\[
\Delta S_i = S_i - f(T_i)
\]

The least square method is again applied to get another polynomial curve equation \( p \)

\[
\Delta S = p(T)
\]

to fit the data points \( (T_i, \Delta S_i) \). For each temperature datum \( T_i \), the salinity upper and lower bounds are
got from the following function:

$$S_{\text{upper}} = f(T) + g_1 \cdot p_1 (T)$$  \hspace{1cm} (5)$$

$$S_{\text{lower}} = f(T) - g_2 \cdot p_2 (T)$$  \hspace{1cm} (6)$$

$S_{\text{upper}}$ is the salinity upper bound versus temperature, in which $g_i$ is the weight coefficient.

In the same way, the salinity lower bound $S_{\text{lower}}$ is got. As this model will be used for the quality control of data automatically, the weight coefficients are expanded empirically (Fig. 5).

![Fig. 5](image)

**Fig. 5** Temperature and salinity data and the corresponding model in 24° - 26° N, 136° - 138° E

1.2.4 Establishing T-S model for the equatorial area and the high latitude area respectively

In high latitudes, when the temperature is between 0 °C and 2 °C, the salinity varies from 31.5 to 34, while in other temperature sections the range is relatively small. At the same time, the salinity data in the temperature section 0 °C - 2 °C, come from different water masses, and do not vary continuously, so the polynomial curve cant be applied directly (Fig. 6).

![Fig. 6](image)

**Fig. 6** Temperature and salinity data and the corresponding model in 44° - 46° N, 176° - 178° E

In the equatorial area, the salinity in the temperature section 15.5 °C - 23.5 °C doesn't vary continuously either. (Fig. 7)

In these areas, the data are divided into two sections. For example, in the equatorial area, the data are divided into two sections, section 1: the temperature below 15.3 °C; section 2: the temperature above...
15.3 °C. In the section 1, temperature is taken as the independent variable, and salinity as the attributive variable. Then the first part of T-S model is got by the method described in Section 1.2.2. In the section 2, salinity is taken as the independent variable, temperature as the attributive variable and the second part was got, then the two models are combined with the right model in this area (Fig. 8).

![Fig. 7 Temperature and salinity data and the corresponding model in 14° - 16° N, 116° - 118° E](image1)

![Fig. 8 Northwest Pacific 2° × 2° T-S model](image2)

2 Quality control of Argo data by T-S model

Compared with the result of the internationally recommended real-time QC, besides remarkably abnormal data (Fig. 9), some amphibolous data may be picked up (Fig. 10). Salinity measurements of ARGO float may experience sensor drifts owing to biofouling and a variety of other problems. This case
can't be picked up by the regular methods, such as range check, gradient check. As the temperature and salinity relationship is more stable in the ocean bottom, the T-S models can pick up these data (Fig. 10).

![Fig. 9](image1.png)  
**Fig. 9** The 91th cycle of float 2900055 temperature and salinity data, compared with the corresponding T-S model

![Fig. 10](image2.png)  
**(a)** 24th cycle of 21852  
**(b)** The 1th cycle of 5900319  
**Fig. 10** Temperature and salinity data, compared with the corresponding T-S model

Because the T-S models are the climatological result based on the historical dataset, they can't detect some noise data effectively (Fig. 11). Moreover, because the spatial and temporal coverage of historical data is not dense enough, the scope is relatively small in some areas. ARGO data in good quality have been selected into the dataset and the model is rebuilt again (Fig. 12).

At present, besides the real-time quality control method recommended by the international ARGO program, T-S models are used in the manual quality control system. We check the ARGO data, and at the same time we modify the T-S models. Later, the T-S models will be used in the automatic quality control.

Until now there are 9 413 profiles, 588 802 stations, and Argo data in Northwest Pacific which have been detected by T-S models (Table 1).
(a) Temperature and salinity profile data   
(b) Compared with the corresponding T-S model

**Fig. 11** The 11th cycle of float 5900227

(a) The 74th cycle of float 2900175  
(b) The 60th cycle of float 4900162

**Fig. 12** Temperature and salinity data, compared with the corresponding T-S model

<table>
<thead>
<tr>
<th>Tab. 1</th>
<th>The result of data being tested by T-S models in north west Pacific</th>
</tr>
</thead>
<tbody>
<tr>
<td>The sum of the data detected</td>
<td>Error data detected by real-time QC(^1)</td>
</tr>
<tr>
<td>566662 data /9162 profiles</td>
<td>24855 data</td>
</tr>
</tbody>
</table>

Note: Real-time QC is offered by the global Argo program

The result shows in the 19031 error data detected by T-S models, about 25 % (4819) can't be detected by real-time QC. Because the two methods are so different, manual check is done. The final result shows that in the error data detected by T-S models, about 50 % can't be detected by real-time QC,
about 25% is also detected by real-time QC, and the rest 25% may be erroneously detected. At present, we have adopted new temperature and salinity data to improve the T-S models. The result shows that the T-S model is more accurate than real-time QC and it is an effective complement to the QC method.

For the error data detected by T-S models, because salinity data are calculated by the temperature, pressure and conductivity data, the salinity data are taken as wrong data first. Then combined with other QC results, the temperature data are checked. For the error data, if they are obviously wrong, we will delete the data directly; if they are constant profile data, the quality control flags are added; if they are the conductivity sensor drift data, the Annie Wong method is adopted to correct the error data, and then the corrected data will be provided to users.

3 Discussion

In general, the T-S model is a useful tool of quality control for the temperature and salinity profile data. But this method is based on the historical data, and the number and the density of the data are critical to the models. The certainty of the estimation will depend on the spatial and temporal variability in the region. We have selected the lately the temperature and salinity profile data to modify the model.

Next, we will set up a new type of T-S model considering the spatial and temporal variations of water-mass structures.

References