International Council for the
Exploration of the Sea

C.M. 1980 / D : 14
Statistics Committee
Ref.: Hydrography Committee

Techniques for reducing large volumes of oceanographic data

by

D. P. Kohnke*

Abstract

Various possibilities of reducing large volumes of data are presented. The advantages and disadvantages of these methods are discussed briefly. A method suitable for the reduction of profile measurements and by means of which the information loss can be kept to a minimum is described in much greater detail.

Résumé

Le présent article décrit les nombreuses possibilités de réduire de grandes quantités de données. Un résumé des avantages et désavantages de ces méthodes est présenté. Une méthode pour réduire les mesures de profils par laquelle les pertes d'informations peuvent être minimisées est traitée en détail.

* Deutsches Ozeanographisches Datenzentrum
Postfach 220, D - 2000 Hamburg 4

1 With modern high-frequency measuring methods in oceanography, the oceanographic data bases are faced with the question whether it is sensible to store every individual item of data or to reduce the amount of data in accordance with international criteria standardised as far as possible.

The problem of dealing with large volumes of data in data centres is primarily not so much technical as philosophical. Before a decision is made on the methods according to which the volumes of data are to be reduced, it must be determined which demands the product, i.e. the reduced data set, must satisfy in order to be sensibly applicable from a scientific point of view. It is precisely here that the difficulty lies, for a reduced data set, which may be adequate for climatological investigations, will not necessarily be suitable for a scientific special investigation.

The question of the limit to which the reduction of the data is made cannot be clearly answered. The answer depends on the later use of the reduced data set.

When coming to a decision on this matter, it therefore seems reasonable to rely on experience gained from the retrieval behaviour of secondary users during the last 10 years. Experience in recent years shows that the data archived in the data centres are utilised almost exclusively for scientific statistical work. Only rarely are the original data retrieved from the centres for special scientific investigations. A significant change in this behaviour of secondary users of data is not to be reckoned with in future. Persons requiring data for special investigations will also in future prefer to take measurements themselves and not rely on the measurements taken by other persons.

2 The aim of reducing volumes of data is:

a) to reduce the quantity of original data to a meaningful minimum amount;

b) to decrease the scientific content of the original data either not at all or only as little as possible;

c) to be aware of the exact information loss connected with the reduction of data and to standardise this loss.

3 For the reduction of volumes of oceanographic data the following types of data are mainly taken into consideration:

a) one-dimensional data arrays (vertical and horizontal profiles and time series);

b) two-dimensional data arrays (e.g. horizontal distributions);

c) possibly combinations of a) and b).

Within the one-dimensional data arrays the profiles and time series measurements form two different groups, the reduced data sets from which are utilised in completely different ways. The criteria governing a reduction of data for these two types of measurements will necessarily differ.

4 Conceivable reduction methods are:

a) methods by means of which a significant subset of data is selected from the original volume of data;

b) spectral analysis with subsequent digitisation of the spectrum;

c) approximation of the data by functions (splines, orthogonal functions) and storing of the coefficients.

5.1 The use of spline functions or orthogonal functions in the reduction of profile measurements is theoretically conceivable, but in most cases not useful. The author is unaware of any literature in which this kind of function has been utilised to reduce large volumes of data.

The disadvantages of this method are also clear:

a) the use of spline functions in the routine reduction of data is in practical terms impossible. Optimisation of data approximation depends on the particular set of data.

b) the number of resulting coefficients (which are then to be stored) may become larger than the volume of original data.

c) the information loss after reduction will vary considerably from data set to data set. The desired standardisation of the information loss can hardly be achieved when using splines.

On the other hand, splines are already in occasional use today for the reduction of long time series measurements (tides, currents). The original measurements are approximated by means of a spline function, and half-hourly or hourly values, for example, are subsequently calculated with this function. With the reduced data set produced in this way it is then of course no

longer possible to analyse more frequent fluctuations or sudden occurrences.

5.2 A further method of reducing time series measurements is spectral analysis. The energy spectrum is first of all calculated from the original data and is subsequently digitised.

This was the method used for example for the international exchange of wave data collected during GATE (the GARP Atlantic Tropical Experiment). The energy spectrum was digitised at 129 points between 0 and 1 c/s. In this way an original volume of approximately 3600 items of data could be reduced to 129 spectral values.

The disadvantages of this statistical method are as follows:

a) phase information is lost;

b) the original data series cannot be reconstructed;

c) it is not possible to make statements concerning extreme wave heights.

5.3 The results of a reduction method in which a significant subset of data is selected from the original volume of data are detailed below. Until now, only vertical profiles but no time series have been reduced by means of this method. Although technically feasible, it is recommended to reduce time series measurements in accordance with different criteria. Statistical values such as mean values and extreme values or equidistantly spaced values (for example hourly mean values) are required most frequently. Statistical values of this kind are not provided by the reduction method described below.

An attempt has been made to develop a reduction method with which the aims described in section 2 above are achieved to a large extent. The original data are not altered. A significant subset of data is selected from the total volume of original data $\{(z_i, x_i)\}$ or $\{(z_i, x_i, y_i)\}$. For this purpose the value x'_{i+1} for z_{i+1} is calculated from the values (z_i, x_i) and (z_{i+2}, x_{i+2}) by means of linear interpolation, and subsequently the difference $\Delta_{i+1} = |x_{i+1} - x'_{i+1}|$ is determined. If Δ_{i+1} is larger than a preset (freely selectable) threshold value L, then x_{i+1} is significant and becomes an element in the reduced volume of data. If on the other hand $\Delta_{i+1} \leq L$, x_{i+1} is ignored.

With a steady slow change of gradient of the curve the calculation of Δ_{i+1} alone is not sufficient. Therefore, a linear interpolation is also made between the final value established as significant, which may be taken as (z_i, x_i), and (z_{i+3}, x_{i+3}), and the deviation $\Delta_{i+2} = |x_{i+2} - x'_{i+2}|$ in the depth z_{i+2} is calculated. If $\Delta_{i+2} > L$, the point (z_{i+2}, x_{i+2}) belongs to the significant subset (reduced data set). If $\Delta_{i+2} \leq L$, a linear interpolation is made between (z_i, x_i) and (z_{i+4}, x_{i+4}). This procedure is repeated until a significant value is determined. This value then assumes the previous function of the point (z_i, x_i) as initial value for the following interpolations.

The disadvantage of this method is that spikes in particular are declared significant and classified as elements of the reduced data set. A clear example of this is given in Fig. 3, which shows on the left a temperature profile and on the right the relevant salinity profile of a bathysonde measurement. The threshold values chosen for the reduction were 0.03 °C for temperature and 0.04 °/oo for salinity (ICES standard). The original data set consisted at this station of 744 value triplets (z, T, S). By means of the above threshold

values the number of triplets was reduced to 198. Each left-hand curve was drawn with the original number of values (= 744), and the corresponding curve with the reduced data set (= 198 values) was drawn 1 cm to the right.

In order to reduce the noise of a data set, this can be smoothed, before reducing the volume of data, by means of a running mean. Hydrographic structures, such as strong vertical gradients, are of course changed to some extent during this process. Fig. 4 shows the same STD station as Fig. 3, but in Fig. 4 the original curve was smoothed with a running mean using 5 successive values, and only then reduced. As a result, the original volume of 744 triplets was reduced to 125 triplets.

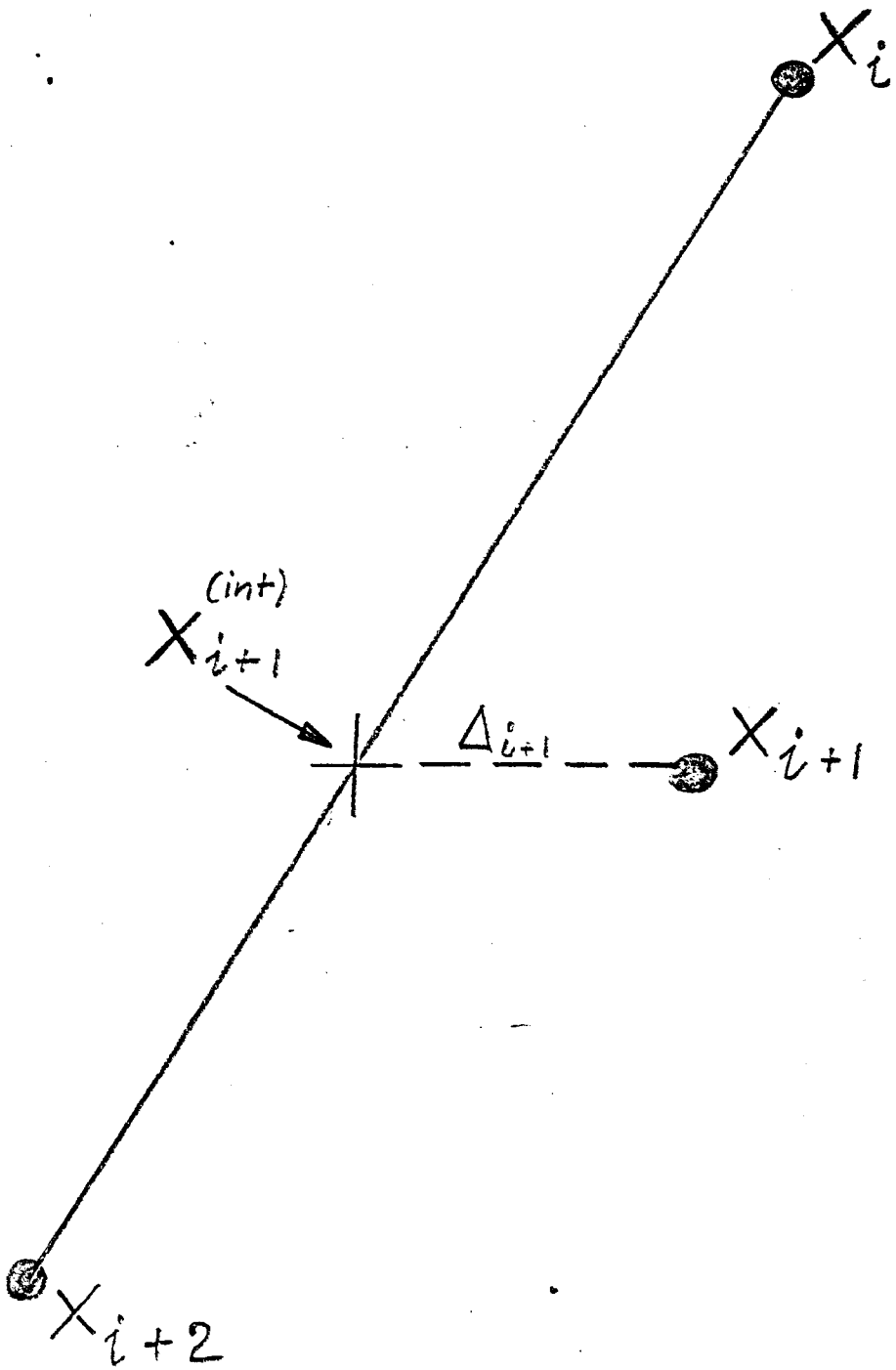
Fig. 5, in which a running mean using 9 successive values was first taken, shows clearly that no significant improvement in the reduction of the volume of data was achieved. It can be seen that the spikes, which are a disturbing factor with this reduction method, disappear even with a running mean using a low number of values.

Fig. 6 shows the same STD profile as in the previous Figures. The volume of data was however reduced with the threshold values 0.1 ^{o}C for temperature and 0.1 $^{o}/_{oo}$ for salinity.

6 Summary

On account of the differing nature of oceanographic measurements and their fields of application, it does not seem reasonable to use a standardised method for reducing large volumes of data. Profile data and time series measurements are fundamentally to be treated

differently. Whereas standard threshold values were set several years ago by ICES for the reduction of profile measurements (0.03 °C for temperature and 0.04 °/oo for salinity) and have been applied successfully by the author in the reduction method he has developed, there are considerable differences of opinion concerning the methods to be applied with the time series. The reason for this is mainly the differing assessments of the purpose of the reduced volumes of data. Every effort should be made to reach agreement on the different points of view in this matter as soon as possible.



$$\Delta_{i+1} = \left| X_{i+1} - X_{i+1}^{(int)} \right|$$

Fig. 1

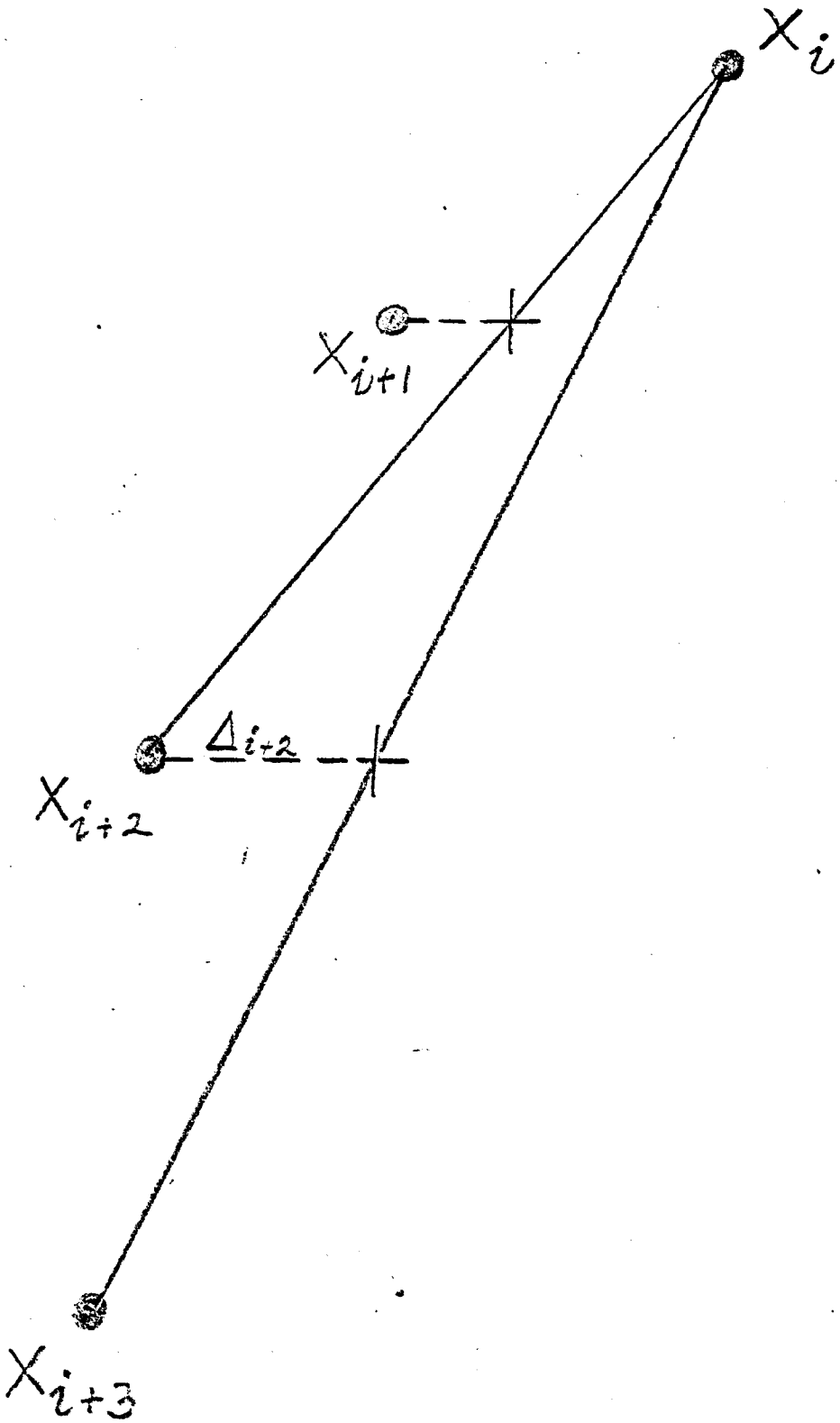


Fig: 2

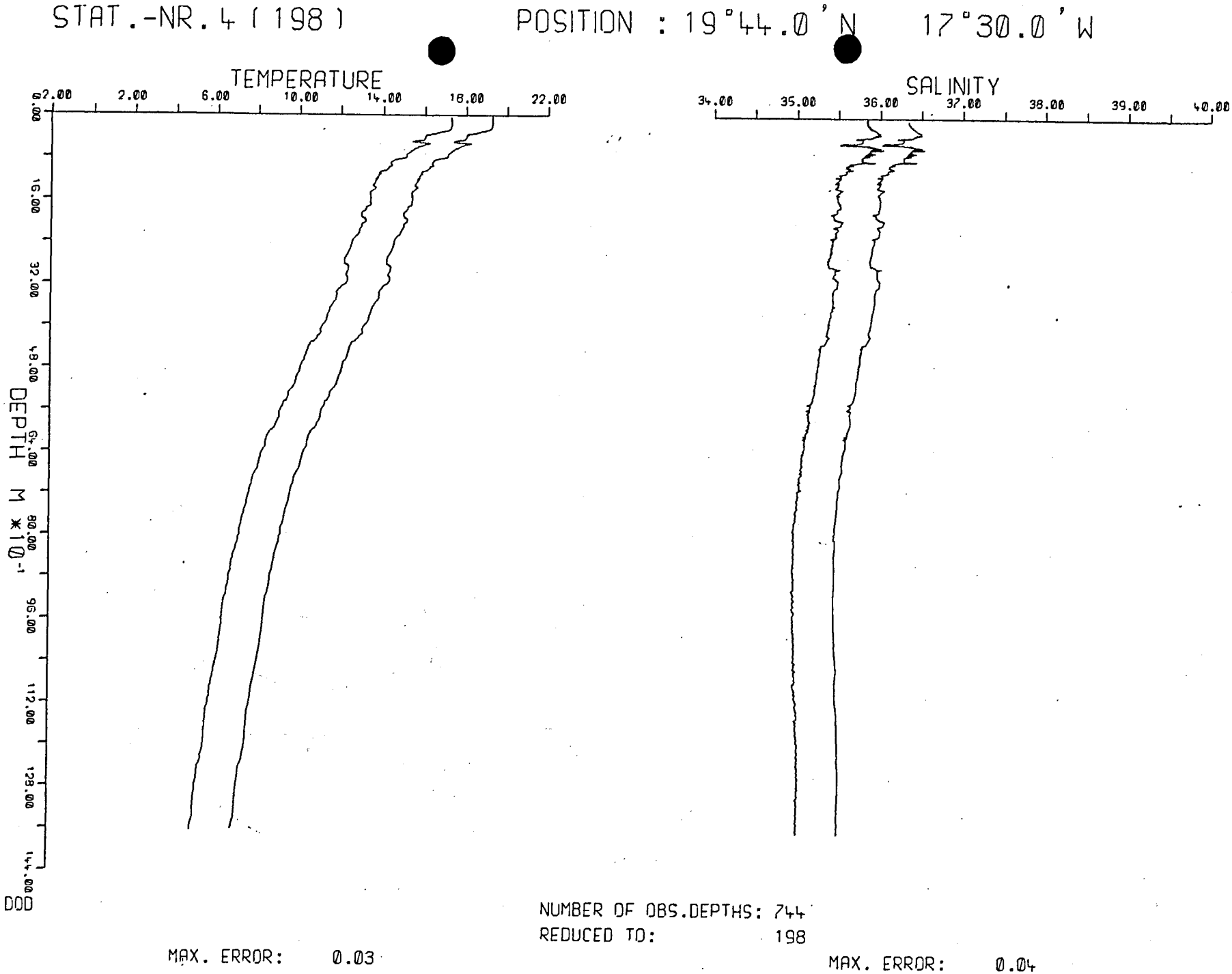


Figure 3 : Corresponding vertical profiles of temperature and salinity (no running mean)
Curves on the left drawn with 744 data points;
Curves 1 cm to the right drawn with 198 data points.

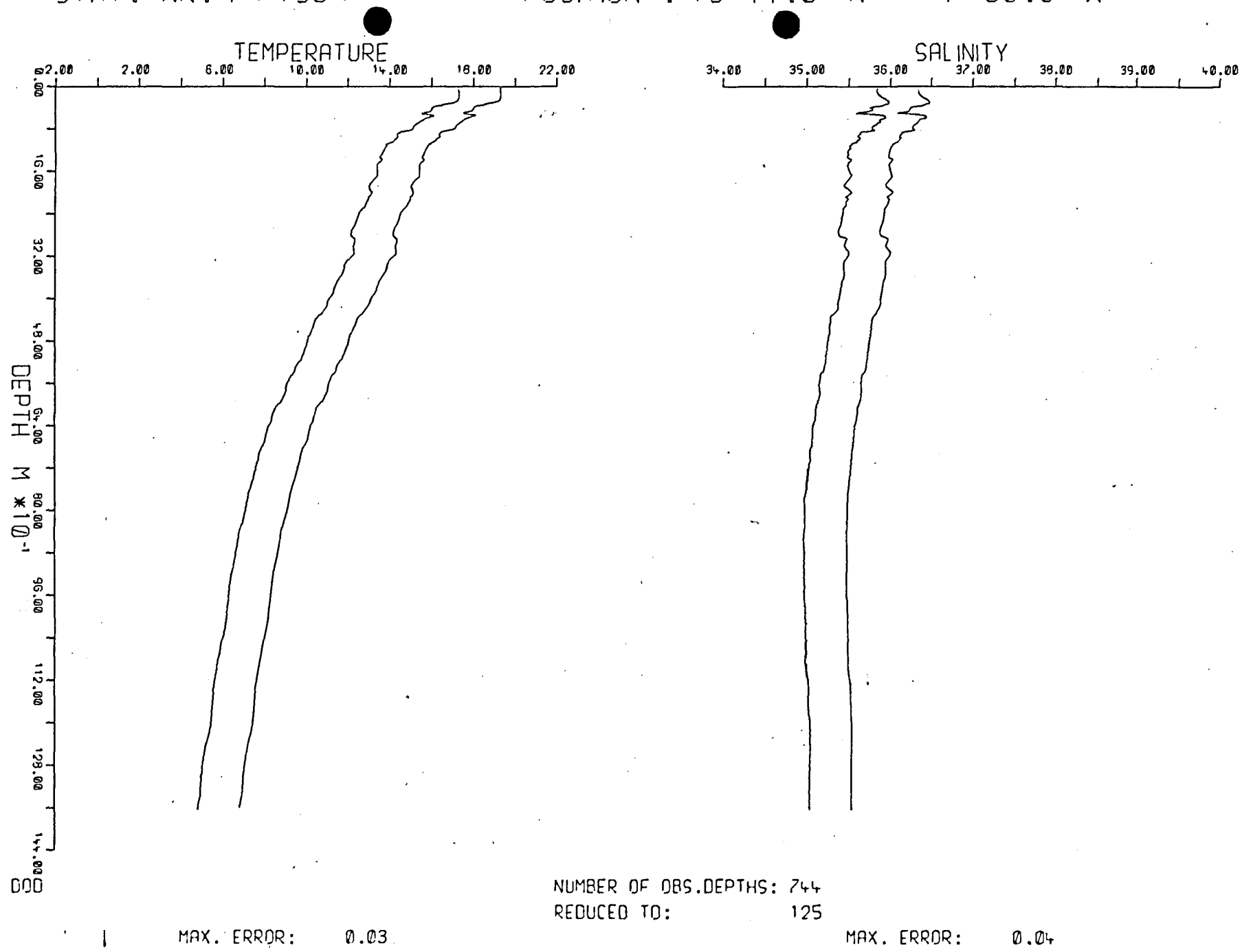


Figure 4 : Corresponding vertical profiles of temperature and salinity (running mean using 5 successive values)
Curves on the left drawn with 744 data points;
Curves 1 cm to the right drawn with 125 data points.

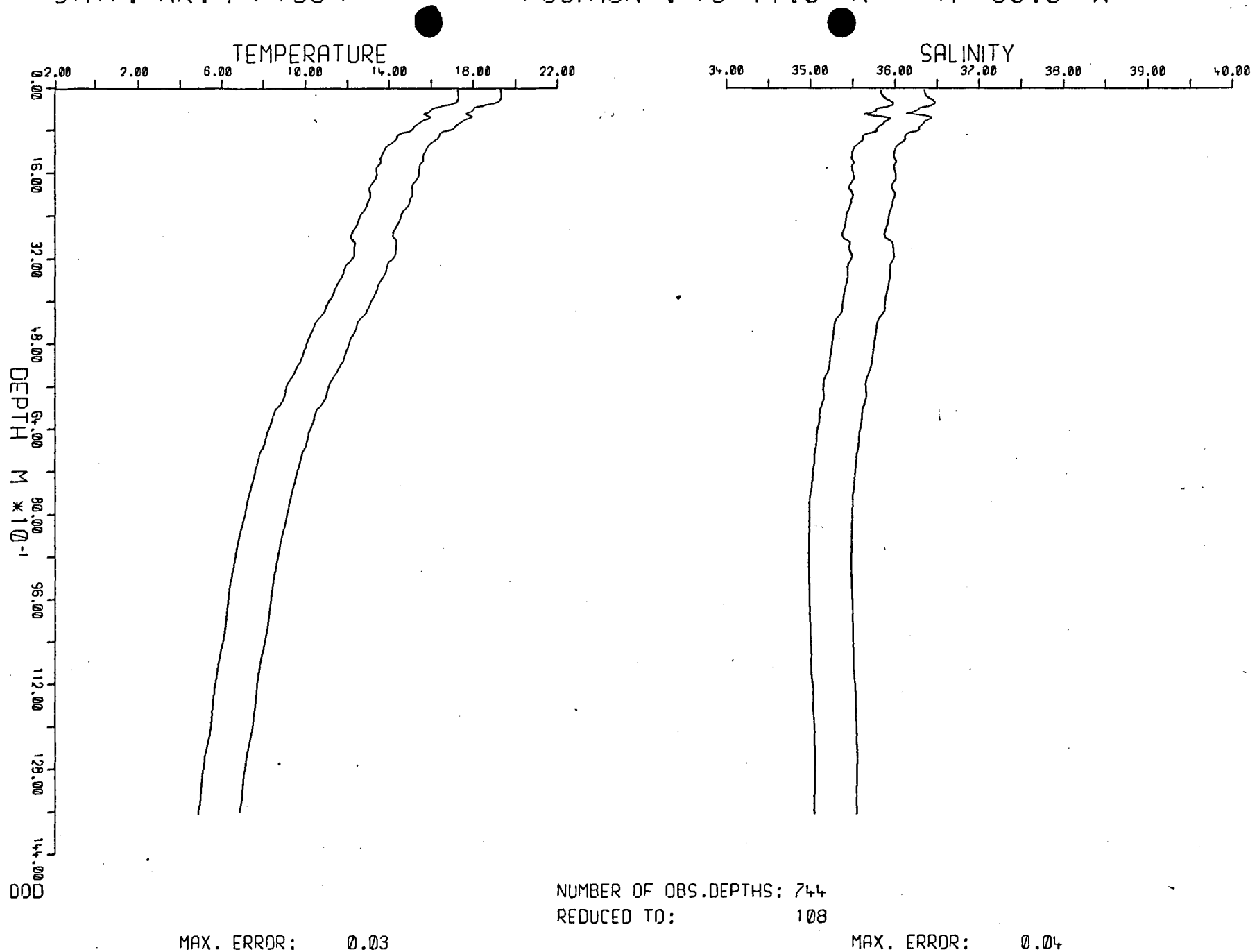


Figure 5 : Corresponding vertical profiles of temperature and salinity (running mean using 9 successive values)
Curves on the left drawn with 744 data points;
Cureves 1 cm to the right drawn with 108 data points.