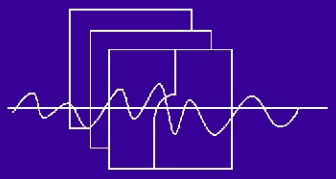
Vol 15 No 4 - December 2004

SADCO SADSO



Southern African Data Centre for Oceanography
P O Box 320, Stellenbosch 7599
South Africa

Email: mgrundli@csir.co.za

Website: http://sadco.csir.co.za/

SADCO is sponsored by ...

Department of Environmental Affairs & Tourism
SA Navy
CSIR Environmentek
NRF (SA Universities)
Namibian Ministry for Fisheries & Marine Resources

Quality control of data in SADCO

The manager of a major European Data Centre once remarked that a data centre will be known for the quality of its data. “If the data centre's data quality is good, there will be many users; if the quality is suspect, they will stay away”.

Because of the universal importance of QC (quality control) global checking guidelines have been developed over the years. While the insight of the data collector/donor is the primary and most reliable screening mechanism, data centres often receive data from sources that do not have/apply this checking facility. They therefore have to construct additional tests to ensure that the data quality is properly controlled.

Previous Newsletters have reported on design aspects of this process, the first one in September 2002. In December 2003 we indicated our planned use of the World Ocean Atlas 2001 (mean envelopes of vertical profiles) to help identify outliers and anomalies in vertical profiles. The July 2004 issue indicated some of the errors that may be anticipated. The latter acted as hypotheses to design and implement the subsequent checking protocols and procedures. The present Newsletter and the next one provides an update on the progress thus far.

It should also be noted that the process of cleaning up the VOS (Voluntary Observing Ships) database has been completed successfully (see article in Newsletter of July 2004). The VOS checks have been incorporated into the VOS loading software, so that the screening process is now ongoing. This is a major step forward in the whole quality control process.

ISSN 1017-2351

Marine Data Quality Control: Ursula pays a visit to the Bedford Institute for Oceanography, Canada

Ursula von St Ange attended an OBIS meeting at the Bedford Institute for Oceanography (BIO) in Dartmouth, Canada in September this year (see previous Newsletter). While she was there, she took the opportunity to meet some of the BIO staff that deal with marine data quality control (QC) issues. She met with Pierre Clement, the manager of BIO's BioChem database, and Mary Kennedy, who looks after the plankton data.

They, like SADCO, rely heavily on their contributors to do the main QC on the data. They do, however, test for outliers by comparing the data against minimum and maximum values, and the records are flagged if their values fall outside the limits. These limits are spatially and temporally independent, i.e. there is only one value, irrespective of where the station is situated, and in which season the station was occupied.

All their data is submitted to the Marine Environmental Data Service (MEDS) (http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Home_e.htm), which is a branch of Canada's federal Department of Fisheries and Oceans (DFO).

MEDS's mandate is very similar to SADCO's, namely to manage and archive ocean data, and to disseminate data, data products, and services to the marine community. The BioChem database is one of many databases managed by them. MEDS use the same QC system as the World Ocean Database (WOD).

Pierre Clements supplied Ursula with some documents from Laurie Devine for QC of bottle data, where use is also made of the WOD ranges for different ocean regions but without temporal (e.g. seasonal) variations. SADCO is also following a similar procedure.

Ursula von St Ange

SADCO's control of the quality of marine data (Part I)

Introduction

As indicated on the front page, the process of cleaning the marine (hydrographic) data has been moving ahead steadily for at least 2 years. The present article is an update on the progress thus far. It will be seen that the past year has been a period of software design and construction, and that SADCO has now virtually reached a stage where the procedures can be universally implemented (checking of all 200 000+ stations in its data base). *All software mentioned below has been written by Ursula von St Ange, including the graphic displays*.

Some philosophical concepts

One of the most important aims of a data centre in its early years is to collate as much data as possible (irrespective of type, quality, format, origin), and this remains important even in mature data centres. Reasons for this include:

a) Often, data centres originate in a rather uncoordinated fashion, with no or few specifications, and data is hoarded as fast as possible (normally there is a massive backlog). After some years, when data holdings have grown, or the funding reduces, rationality starts to prevail and limits are introduced, mostly in terms of geographic dimensions and data types.

b) The aim of the data centre is to serve its users, and because of the often diverse user community the data centre needs to have data virtually everywhere and for all times, just in case somebody asks for it.

c) A data centre can have other obligations, and in some cases (especially in developing countries) this includes repatriating data collected in its territorial waters or EEZ.

However, as the data centre progresses, almost like Mazlow's Hierarchy of Needs, the “needs” (=role) of a data centre tends to start including other aspects of data management. One of the main components here is the aspect of data quality. If the data centre is a national facility, this stage may tend to parallel a country's intrinsic economic and social development, or the need to differentiate between users demanding quality rather than quantity (e.g. users wanting WOCE quality data), or just the desire to ensure that the quality of the data in a data centre is as good as possible.

SADCO, like most data centres, relies on the data donors to ensure the data quality. Only the processors of data have access to metadata like calibration figures, log books, or even the insight of having collected the data themselves.

Implementation of the QC procedure

The quality control procedure tends to exclude calibration issues (which are assumed to have been attended to by the data processor/collector), but would include testing for spikes in the vertical profiles, positional errors, time errors, etc. Some errors may be of such a nature and scope that they would require more time and closer insight into

SADCO's control of the quality of marine data (Part I) continue

instrument type than what SADCO would like to invest, and such data could be returned to the data donor.

The checks that have been, or will be, introduced are listed below. They agree with methodology proposed by IOC (Intergovernmental Oceanographic Commission) for data quality control (GTSPP Real-time quality control Manual, Manuals and guides 22, Unesco 1990).

The "starting point" of checking is the assumption that some ("basic") parameters are correct (if all parameters are incorrect, it becomes virtually impossible to check for errors). E.g. position and time are essential, basic parameters that need to be validated before subsurface parameters can be checked. If the position is wrong, it becomes virtually impossible to check the vertical profiles (since one doesn't know where the data was collected).

a) Surface metocean parameters

There are obvious "global" limits for parameters (e.g. a latitude value cannot be larger than 90º, and wind direction not larger than 360º, temperature cannot exceed certain values, etc.)

For the surface parameters of the hydrographic stations SADCO combines the WMO limits with regional and temporal limits for metocean parameters, and a comprehensive list has been provided in the December 2003 issue of the SADCO Newsletter. These were developed from, and have been applied to, the database of VOS (Voluntary Observing Ships) observations during a clean-up process in 2004, and now form part of the routine screening during the loading of new VOS data. They are also applied to surface observations of hydrographic data.

b) Station positional check

The station position is of cardinal importance to the information contained in a hydrographic station. Typical errors include: swopping of the latitude and longitude; punching/editing errors such as the latitude or longitude has been truncated (3º in stead of 30º) or the minutes have been left out, hemisphere is wrong; date/time is wrong, etc. SADCO checks the position in the following ways:

- *Global checks*: the position must be valid, namely that the latitude must lie between +90ºand -90º, and longitude between +180ºand - 180º
- Checking whether the position lies on the correct side of the coastline (the so-called *overland check*). SADCO applies this check only for the VOS data, not for the marine data. The reason for this is that VOS data is mainly deep-sea data, and the distinction of whether the reported position is overland or not, is clear. In marine data, where positions can be close to the coast, even on the beach or in estuaries or river mouths, the

land-sea boundary lacks sufficient spatial resolution and accuracy to avoid ambiguity.

- Checking whether the indicated *bottom depth* agrees more or less with an objective estimate of the bottom depth. Here, SADCO uses the Etopo2 data set (2-minute Gridded Global Relief Data), a database of global bottom topographies in 2' x 2' blocks, generated by NOAA National Geophysical Data Centre. Significant differences can nevertheless be expected close to the coast or in the vicinity of the shelf edge (or other bottom topographic features).
- *Ship speed check*: If the ship's expected speed between stations, as computed from the various stations' times and positions, is exceeded by a significant margin for one station, the position of that station needs to be verified.
- *Visual checks*: It is sometimes easy to identify anomalies and outliers if the cruise station pattern has a regular grid.

c) Bottom depth check

Apart from using the bottom depth as a positional verification, the bottom depth has to be correct in itself. The test here is

- a comparison with an independent data source, and SADCO uses the Etopo2 database for this;
- comparison with the bottom depths of other stations in the vicinity
- verification that the recorded bottom depth is larger than the depth of the deepest sample.

Planned further developments

The checks mentioned above are not being implemented yet. Before this part of the process is tackled, the following aspects still need attention:

- Additional graphic plots will be considered (e.g. a T/S plot could assist with identifying anomalies).
- The outcome of the checking procedure will need to be decided. Should suspect data simply be flagged? Should editing be done? When will data be returned to the donor (this will not be possible for historic data)?
- The process of scanning the whole data base will be undertaken, and this is quite a major task that needs to be planned carefully.

The checks will be built into the loading programme, so that the data screening can be done in real time in future.

Output of the QC process

The QC process has 3 stages:

- Apply software checks to selected data sets.
- Visually display the data and the indicated error, to assist the data checker.
- Implement an edit protocol in the form of a quality flag, or error correction where the correct value is obvious.

SADCO's control of the quality of marine data (Part I) continue

Examples of the various checks

The checking software is run on a selected data set (e.g. a cruise), and a print is obtained of

- All stations and positions
- Data types contained in the station
- Vessel speed between stations, with those speeds above 20 knots marked as possible errors
- Stations that seem to be overland according to Etopo2
- Stations north of 10º N (SADCO's northern limit)

The graphic support to investigate possible errors emerging from this process is in the form of various plots:

Louise Watt, who is a seasoned data editor, called the display of the station positions (Figs. 1, 2a and 2b) "brilliant" as an aid for data editing/ checking.

The next Newsletter will provide insight into the checking of vertical profiles, and examples of the corresponding plots.

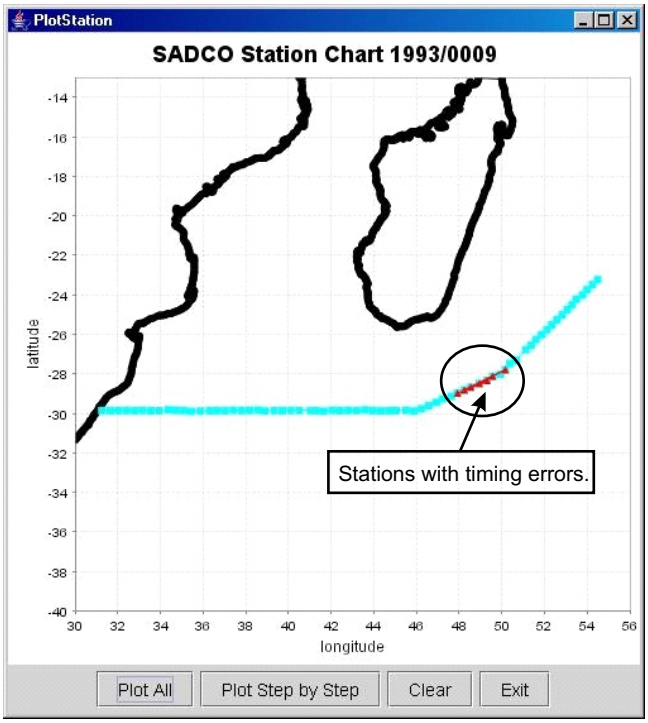
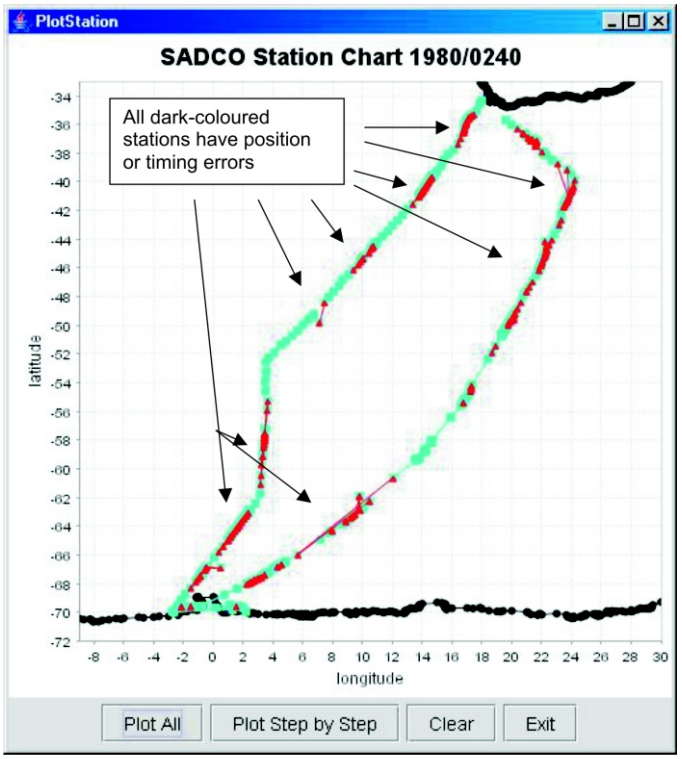
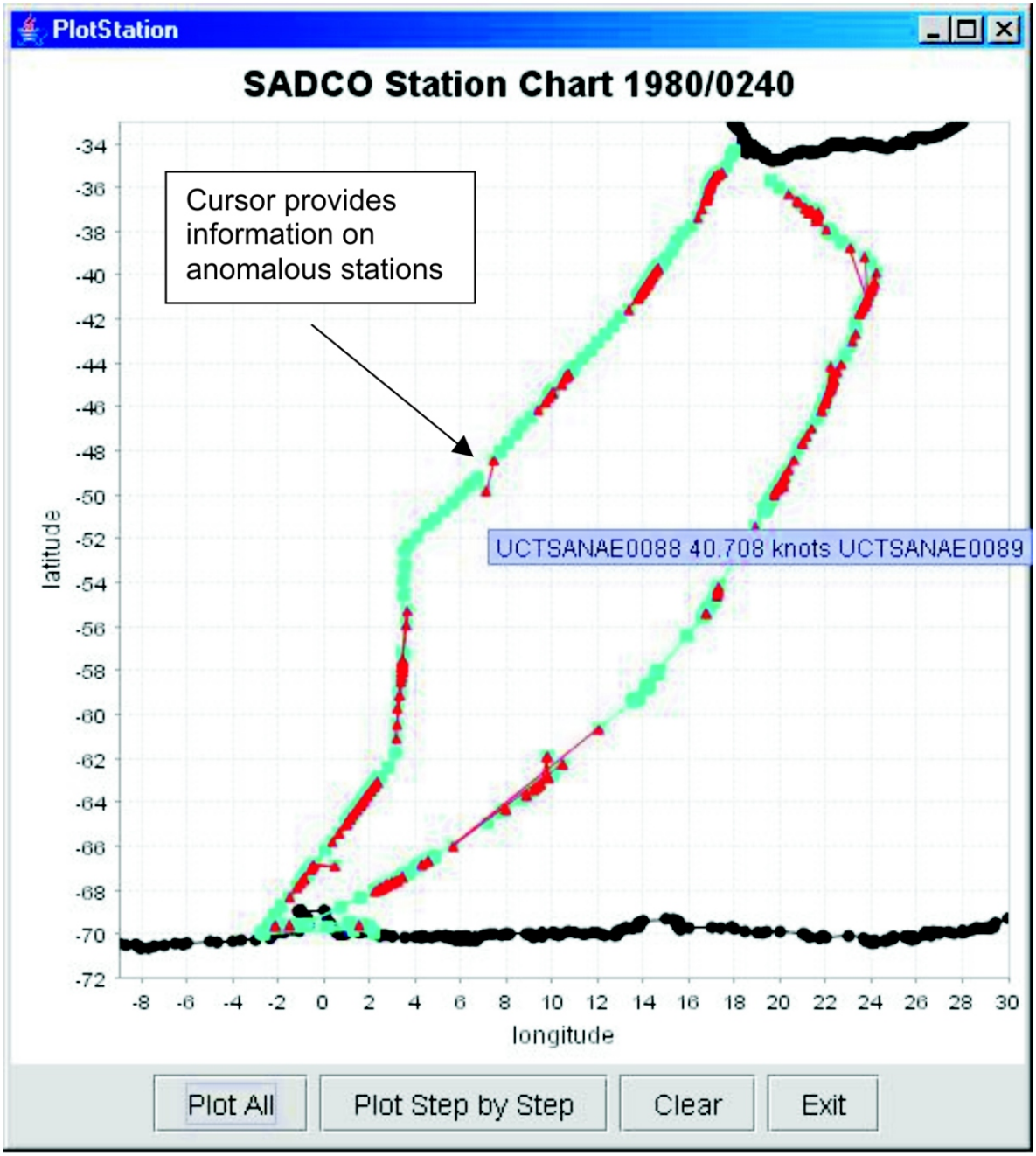


Fig. 1. Track plot of a cruise between Durban and Mauritius. All stations visually appear to be correctly located, but because of timing errors, some stations have produced too high vessel speeds. These stations are coloured differently (on the original colour screen).

Fig. 2 Track plot of two transects between Cape Town in the north and Antarctica in the south. The stations in darker shades (originally red) indicate vessel speeds exceeding the threshold of 20 knots. The cursor identifies non-anomalous stations simply with the StationID (Fig. 2a), while anomalous stations are identified pairwise with the anomalously high speed indicated (Fig. 2b).




PlotStation
SADCO Station Chart 1980/0240
Cursor provides information on anomalous stations
UCTSANAE0088 40.708 knots UCTSANAE0089
latitude
longitude
-34
-36
-38
-40
-42
-44
-46
-48
-50
-52
-54
-56
-58
-60
-62
-64
-66
-68
-70
-72
-8 -6 -4 -2 0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30
Plot All
Plot Step by Step
Clear
Exit