Vol 19 No 4 - December 2008

SADCO SADSO

QUALITY CRITERIA FOR TIME SERIES DATA

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

Manager: Marten Gründlingh

Email: mgrundli@csir.co.za

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

SADCO is sponsored by ...

- Department of Environmental Affairs & Tourism
- SA Navy
- CSIR
- SAEON
- Namibian Ministry for Fisheries & Marine Resources

ISSN 1017-2351

Background and motivation

Up to 1990 SADCO loaded data exactly as it was received from the data provider, under the policy that data quality was the responsibility of the data donor. This meant (and still means) that data had to be checked and validated before submission to SADCO.

From 1990 to the present SADCO has been trying to add value to the data by conferring with the data provider on questionable records, and judiciously editing data where appropriate.

Over this period SADCO's data holdings have increased continuously. In addition, the initial range of data types (hydrographic profiles of temperature and salinity, and the VOS weather observations) has systematically been expanded to include other useful data types (e.g. additional subsurface parameters, wind data from automatic weather stations, current meters, etc). The result is that the data has been growing in both quantity and diversity.

Lately, SADCO has come to realise that the amount of incoming data was, at times, somewhat like a tsunami that tended to overwhelm the capacity of the data centre. The largest single loading events during the past decade comprised the ~20 000 XBTs loaded in 1998; 50 000 foreign stations loaded in 1999; 36 000 stations loaded in 2006 and 22 000 ARGO profiles loaded in 2007. These "data tsunamis" have contributed about 50% of SADCO's present hydrographic data holdings. In addition, the effort needed to deal with local data seems to have increased, probably as a result of organisations losing capacity and expertise, and greater pressure on a smaller number of data collectors. It has therefore become difficult, and on the occasions of the large data "tsunamis, impossible for SADCO to check and correct data all errors.

New process

Between the two options when dealing with incoming data (outright rejection, full acceptance) the need was identified to adopt a compromise approach between rejection and acceptance. This was spawned

by the unfortunate reality that organisations may not have the ability to produce data that is 100% checked, and if questionable data is returned to the data provider the data may never get into a long-term storage facility. (It should also be realised that retrospective improvement of data quality by the data provider after the data has been loaded, is probably not going to happen either).

A decision was therefore made to adopt a system of quality control and assessment, making use of SADCO's oceanographic insight as well as criteria developed internationally, to identify and label various types of anomalies in the data. This meant that data would not be corrected (edited), but that errors would be flagged to alert the eventual analyser to correct for, or exclude them from, a particular analysis.

The same philosophy holds for time series data. The criteria for time series are different to those for profiles, because of different parameters, different measuring methodology, etc.

"Data tsunamis" have contributed about 50% of SADCO's holdings of vertical profile data.

Criteria for quality checks

The following are the types of checks that SADCO will be performing on incoming time series data:

1. **Metadata checks**: [latitude, longitude, institute, data, time]. No flags are allocated here but **location errors** are considered fatal (i.e. no certainty on WHERE the data was located unfortunately makes the data useless) and such data will not be loaded.

2. **Broad range** check. I.e. does the time series as a whole lie between accepted limits?

3. Are there visible **spikes** present in the data? This would include "traditional" spikes (anomalous, singular departures from a reasonably steady-state background), as well as a noticeable degree of **noisiness.**

4. Is there evidence of **possible drift of the data sensor**? It is realised that slow changes in the data level may be confused with normal, seasonal variation in the data, so any drift should be characterised by a long period of steadiness (= good data) followed by a somewhat exponential decrease (e.g. as the battery fails, or, more probably, the sensor becomes fouled – see Fig. 1). Also included in this would be unexpected, step-like changes in the recording level without evidence of corresponding variation in environmental forcing.

5. Are there visible **gaps** in the data (Fig. 2)? Included here would be occasions where one of the data channels ends prematurely (compared to the other channels).

6. Are there visible pre-deployment **leaders** or post-recovery **trailers** (see Fig. 2)?

The mention of "visible" presence of errors is not without reason. Some checks are not easy to be undertaken programmatically, while they are more easily recognised by visual inspection (even at the risk of subjectivity). An exception is the broad range check, where the data can rather easily be compared to agreed limits. *For the present, SADCO is doing all the checks visually, but the best process is a programmatic identification of the errors, followed by a visual validation by an experienced observer*. This is foreseen as a future development.

First useful checks and flags have been formulated for time series data (current meters,...) in SADCO. This will replace the considerable effort required previously to correct errors in incoming data.

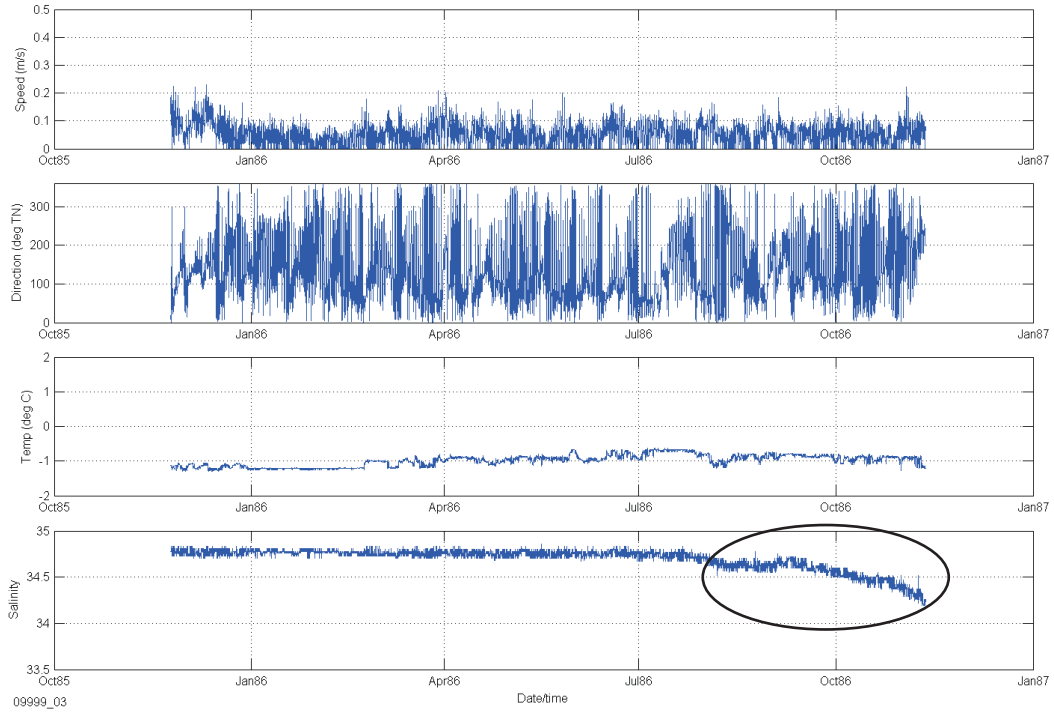


Fig. 1 Time series showing an apparent drift in the salinity sensor (lower panel).

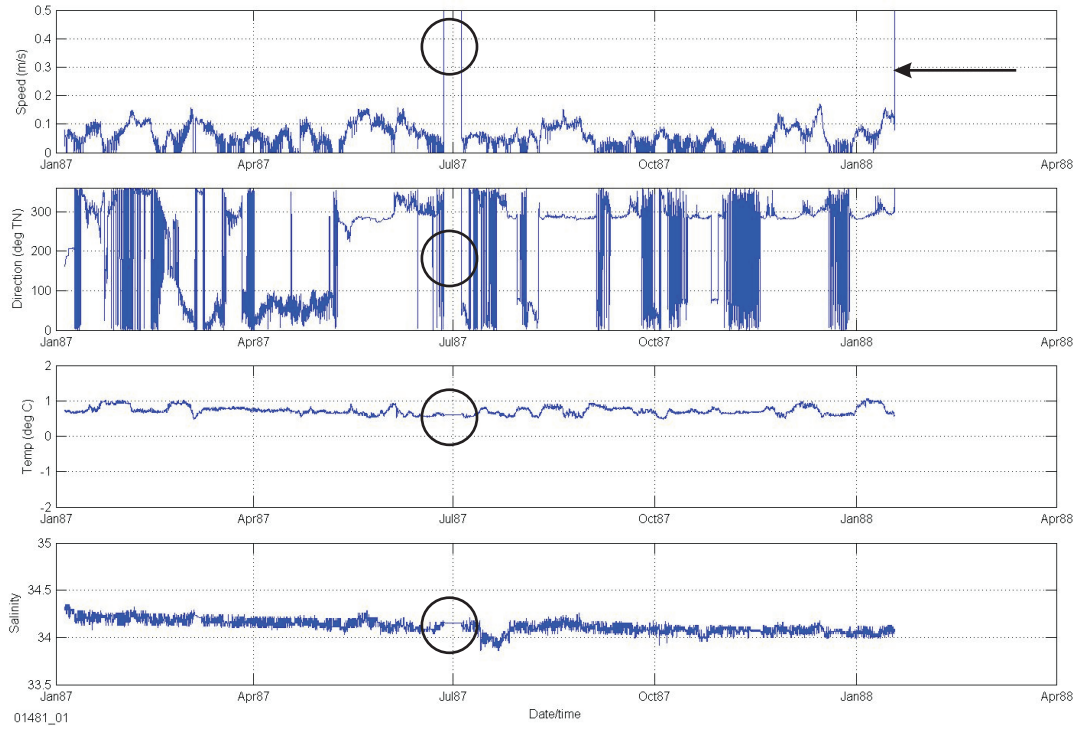


Fig 2. Time series of current speed and direction, with indicated data gaps (circles) and post recovery trailer (arrow).

ANTARCTIC CURRENT METER DATA LOADED

SADCO has recently loaded current meter data in the southern parts of its target area, closer to the Antarctic continent.

The locations of the current meter moorings are indicated in Figure 3, along with other deep-sea moorings.

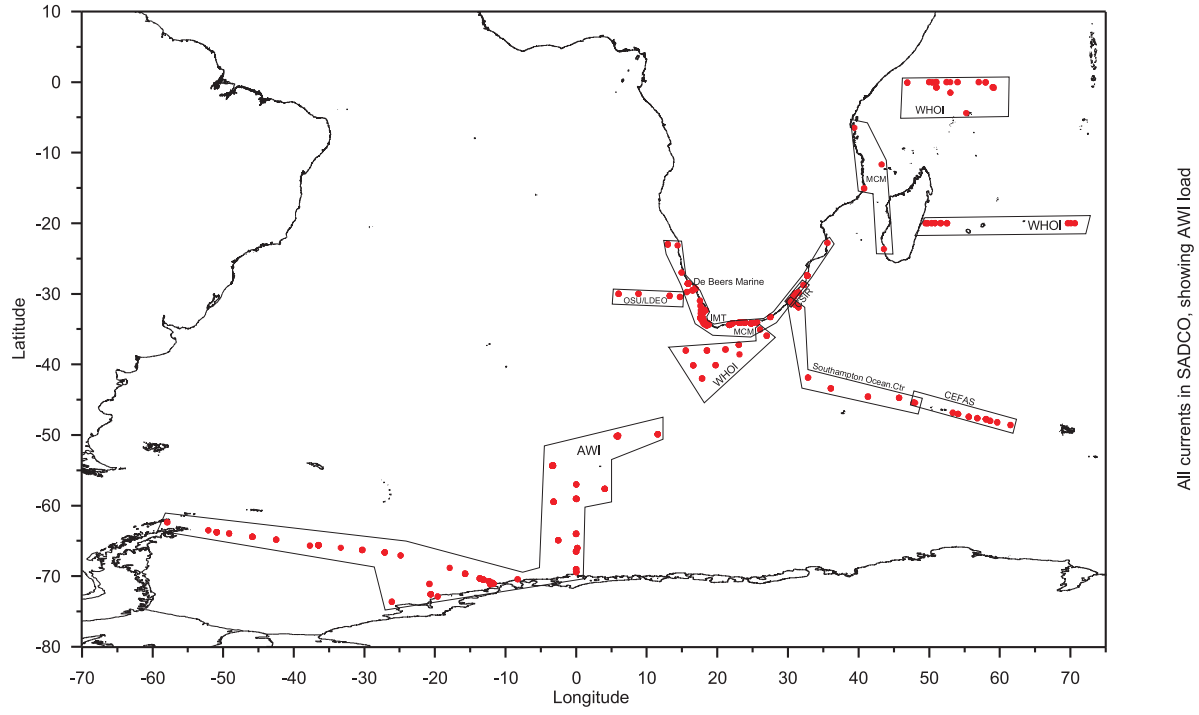
The overall data set is quite huge and contains too many deployments to be listed individually in this Newsletter, but the following information provides some insight into the characteristics of the deployments:

- The data set comprises 220 current meter deployments and 6 thermistor strings. The addition of this data set brings the total number of current meter deployments in SADCO to 1048.
- The data was collected by the Alfred Wegener Institute for Polar and Marine Research (AWI).
- Parameters recorded are current speed, direction, temperature, salinity and pressure. More than 95% of the deployments had speed, direction and temperature.
- The time period covered by the moorings extended from 1985 to 1998.
- Individual moorings were typically 1-2 years duration.
- Time interval of measurement was between 30 and 120 minutes, with the bulk of the instruments recording at either 60 or 120 minutes.
- Instrument depths varied from 30 m to 5 100 m.
- Typical average recorded speed was < 10 cm/s with maximum speeds < 50 cm/s.
- Average temperature was < 1° C, with a minimum of -1.9°C and a maximum of 3.4°C.

The quality of each deployment was visually assessed according to the criteria given in the article in this Newsletter (see p 1 & 2).

The number of deployments (226) make this the single largest loading of moored subsurface instruments in the history of SADCO.

Fig. 3 Locations of all moorings in SADCO. The new data is the group marked AWI.



ASSESSMENT OF SILICATE DATA QUALITY

As part of SADCO's overall quality assessment process reported a year ago (see Newsletter of December 2007) it was found that 31% of the silica data in SADCO is suspect (i.e. had failed a quality check).

On a philosophical note, to include such a data set in a database seems somewhat meaningless. However, discarding data of that nature is not really a practical option. In addition, any re-analysis is out of the question, and feedback to the data provider is not always feasible (a general problem with older data, where the data collector is not available any more, and often even the institute has ceased to exist).

The issue of silicate data quality triggered a number of aspects that needed to be addressed systematically, in order to get the house in order. In the end the picture was much rosier.

Identifying the problem

At the time of designing and applying the quality system (2007), SADCO was also busy loading a large amount of data extracted from the World Ocean Database (WOD) 2005.

Because the incoming WOD silica data was kept in a separate file, the QC criteria were applied specifically to this data set, but not to all the rest of the silica data already in SADCO. The eventual table that was produced to show the outcome (see Newsletter of December 2007) inadvertently omitted to indicate that the results, as far as silica was concerned, referred only to the incoming WOD data (and not all the silica data).

So there were really two issues:

- the reported QC process on silica excluded a large part of the silica data holdings
- the data that was checked seemed to contain an unacceptably large amount of errors.

Rerunning the QC process

Before re-running the process for all the silica data it was decided that the occasion was opportune to revisit the quality criteria (this will probably be done at regular times in future) and make small adjustments. This set of criteria is indicated in the Table below.

QUALITY CHECKS FOR HYDROGRAPHIC STATIONS

Flags for the entire profile	
0	Profile accepted
1	Failed annual standard deviation check
2	Two or more density inversions
3	Failed spike test
Flags on individual observations – depth flags	
0	Accepted value
1	Duplicates or inversions in recorded depth
2	Density inversion
Flags on individual observations – parameter flags	
0	value accepted
1	range outlier (outside of broad range check - per basin)
2	failed inversion check
3	failed gradient check
4	failed spike test
5	failed annual standard deviation check (profile envelope / 5° x 5° blocks)
6	combined gradient and inversion checks
7	flag 1 and flag 2 (broad range & inversion)
8	flag 1 and flag 3 (broad range & gradient)
9	flag 1 and flag 4 (broad range & spike)

If a measurement could be made in the sea with 100% certainty of the outcome, the measurement is actually superfluous. The quality assessment process tries to separate the natural (= environmental) variability from the variability introduced by equipment malfunction or by the operator.

The QC process for silicate has now been re-run on ALL the silicate data in SADCO, using QC envelopes that have been defined more realistically.

Redefining the QC envelopes for silica

In our opinion the "31%" of flagged data mentioned above originated from apparently over-stringent and too close-fitting "envelopes" used to identify outliers. Having envelopes that are too "narrow" tends to defeat the object of only identifying spikes and other anomalies (that can be ascribed to equipment malfunction or operator interference).

To correct for this, it was decided to formulate new envelopes for the silicate data. SADCO inspected all the silicate values in its target area by plotting the depth-silicate profiles, (we have no independent means to assess the spread of the data other than using the data itself). This is shown in Fig. 4 for the three ocean basins around southern Africa.

The graphs show that there is a fair degree of "coherency" in the vertical profiles (just by looking at the clustering of the data points), but that there are also isolated "spikes" or outliers (and even negative values, not shown in the figure).

While the initial flagging used envelopes applicable to 5° x 5° degree blocks, it was considered that the data density at that spatial scale was too low for systematic definition of the envelopes. Three, more "global" areas were therefore chosen (Indian Ocean, South Atlantic Ocean, Southern Ocean) and rough envelopes (not based on the calculated spread of the data) defined for these (shown in the Figure). These envelopes are narrower than the "broad range checks" used to assess the data quality, but wider than the initial 5° x 5° envelopes. In this way, they represent an intermediate step in getting to grips with a sensible process of assessing the silicate quality. It is planned that the envelopes will be redefined again in future, by quantifying the spread at various depths and defining the envelope at that depth as a multiple of the standard deviation. This will make the envelope simulate the depth variation of the profile scatter plot, and be better tuned to the observations (than the visually defined envelopes used now).

a) Outcome of the new quality assessment

The assessment of the data quality has now been (re)done, and the outcome is presented in Table on p8.

Only the updated values for the silicate have been replaced in the original table (the other nutrients will also be systematically handled and updated in the near future).

In terms of the silicate, the following points are noteworthy:

- The **number of stations** with silicate is now 20 783, compared the previous count of 3 026. This correctly represents the amount of silicate data in SADCO. The corresponding **number of observations** has increased from 42 031 to 203 694.
- The number of errors picked up are now an order of magnitude smaller than before, because the envelopes have been widened. It is believed that the present formulation is a better representation of the data quality than before, even though it will be improved in future (as indicated above).

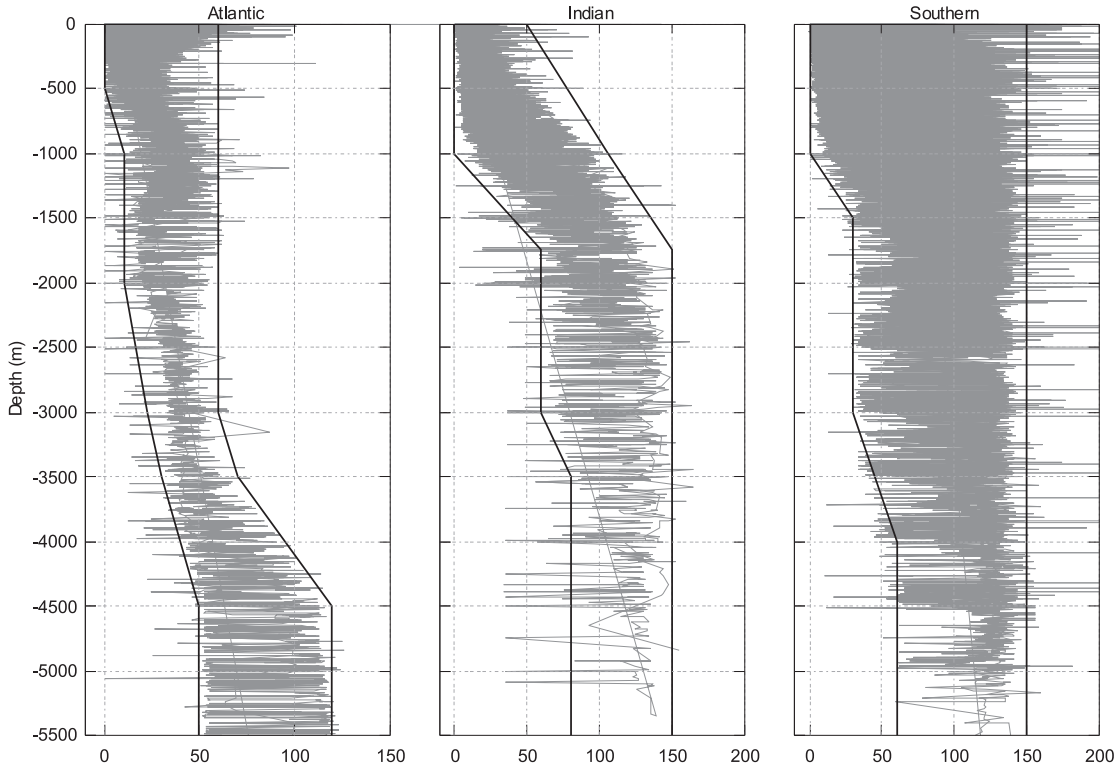


Fig. 4 Plots of all silicate data in three ocean basins in SADCO. Also indicated are the envelopes used in determining outliers in the profiles. "Atlantic" is defined as the SADCO target area west of Africa and up to 34°S, "Indian" as the area east of Africa up to 34°S, and "Southern" as the rest of the target area.

Table 1. Quality assessment of SADCO profile data (*updated with silica only*)

Surveys

No of Surveys	5172

Stations

	Counts	%
No of stations:	243566	100.00%
No of stations failed speed check:	26863	11.03%
No of stations over land:	5406	2.22%
No of stations with no water-physical records:	5276	2.17%

Parameters

Parameter	DPTH	OXY	SAL	TMP	NO3	PO4	SIO3	CHL	DIC	PH
Total stations with parameter		48363	90098	219625	17862	30738	20783	11191	394	4251
% of total stations		19.9%	37.0%	90.2%	7.3%	12.6%	8.5%	4.6%	0.2%	1.7%
Total observations	31870895	6877493	13003406	31807494	145994	289984	203694	127106	1776	52328
% of total depths	100.0%	21.6%	40.8%	99.8%	0.5%	0.9%	0.6%	0.4%	0.0%	0.2%

Flags: Counts

Parameter	DPTH	OXY	SAL	TMP	NO3	PO4	SIO3	CHL	DIC	PH
No of stations with profile flags	0	3051	6289	9521	685	2329	285	227	9	35
No of obs with flags	2262	4858	15336	71988	4489	1010	1249	10465	101	418
No of obs failed inversion/gradient checks	**2262**		**76**	**25083**	**3496**	**3**				**1**
No of obs failed env check	0	140347	62123	371967	3682	11937	1217	1531	9	35
No of obs failed broad check	0	4858	15260	46905	993	1007	231	10465	101	417
No of obs failed spike check	0	264	3503	4112	25	26	47	21	9	35

Flags: Percentages

Flags: Percentages	DPTH	OXY	SAL	TMP	NO3	PO4	SIO3	CHL	DIC	PH
No of stns with profile flags		6.31%	6.98%	4.34%	3.83%	7.58%	1.37%	2.03%	2.28%	0.82%
No of obs failed inversion/gradient checks	0.01%		0.00%	0.08%	2.39%	0.00%				0.00%
No of obs failed env check		2.04%	0.48%	1.17%	2.52%	4.12%	0.60%	1.20%	0.51%	0.07%
No of obs failed broad check		0.07%	0.12%	0.15%	0.01%	0.01%	0.00%	0.15%	0.00%	0.01%
No of obs failed spike check		0.00%	0.03%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%