Vol 13 No 3 September 2002

SADCO SADSO

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

Email: mgrundli@csir.co.za

Website:
http://fred.csir.co.za/ematek/sadco/sadco.html

SADCO is sponsored by ...

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

New "coat" for SADCO *newsletter*

I assume that everybody noticed that the cover of the SADCO newsletter had changed.

The previous cover design (by Monique Knoetze) had been in existence for about 4 years, and in itself had been the 2nd design since the Newsletter started in 1990.

The present cover was designed by Magdel van der Merwe, who is now handling the Newsletter layout.

This is also an appropriate time to introduce Magdel. She is a professional "report producer", which includes typing (very little, she stresses!), report layout, and design, report production, web sites, checking, file conversion, etc. Presently, in a time of many reports (often complex) and publications, and where organisations stress the functional and esthetical appearance of reports, this role is essential.

Welcome Magdel!

The present newsletter focuses on aspects of quality control. This is an extensive subject, one that is of particular relevance to data management. Some future editions of the newsletter will also be devoted to this important subject.

An experienced French data manager once told me "A data centre will be known for the quality of the data it makes available. If you distribute bad data, users will stay away from you". SADCO does not serve only to store the best data, but as a national facility also looks after data collected by Southern African organisations. The quality of such data may have been good at some stage, but has since been superseded by higher quality data. The old data is nevertheless not discarded, because it is believed that environmental data represents a specific capture ("snapshot") of the conditions of that time.

Magdel van der Merwe

ISSN 1017-2351

Checking the quality of VOS data in SADCO (1)

SADCO is keen to have the best quality data in its databases. For this it relies on the supplier to check the quality. This is a time-consuming, specialised work, for which SADCO does not have the funding nor the access to metadata required for proper quality control.

Most requests submitted to SADCO for information based on VOS data, concerns **averages** (mean wave heights, mean temperatures, etc). Most of our products are designed to produces these averages, standard deviations, etc.

Recently, a request was submitted to SADCO to provide data on **maximum** temperatures. Suddenly, upon extraction, some temperatures appeared from the data that were obviously questionable. It seemed that these values were few, and had previously not been recognised because they contributed very little to averages. Only now, when extracted for their individual characteristics, were they recognised as anomalous. A superficial check indicated that such anomalous data exist on the **high** (= maxima) as well as the **low** end (= minima) of the spectrum.

The SADCO Steering Committee agreed that the matter needed attention, and a task was scheduled for the present financial year.

VOS data

Given that VOS data has been collected for almost 150 years, quality control (QC) issues are slightly more complex than for a scientific experiment planned under present conditions. Virtually no metadata in terms of instruments etc are available, nor do we know the "trustworthiness" of the data collector.

VOS data does not seem to be screened very thoroughly before submitted to SADCO for loading. WMO (World Meteorological Organisation) screens (see Table on page 7 of this Newsletter) are in themselves fairly "lenient", but are presently not applied by the supplier. It is left to the individual analyser and user to identify and disregard anomalies.

Because of SADCO's reluctance to discard any data, it was decided that

- Quality control procedures should be confined to eliminating the obvious errors only.
- The process and outcome of applying QC procedures will be discussed with researchers and users in the field. **This article is the first in a series to invite such feedback.**
- Whatever criteria are applied, no data will be discarded. Suspect data would be flagged or simply transferred to another file, which would keep it accessible (maybe not to the routine user) for future reference, if required. At the same time, these values would not contaminate any averages or other statistics derived from the data.

To establish QC procedures that would have benefit for SADCO users, we investigated data distributions in the following way:

- VOS data was extracted between 0° and 50°E (the VOS "Main" data base), and from 1960 onwards.
- The extracted data was split between a "west" segment and an "east" segment (on either side of the continent). South of the continent the dividing line was 25° East.
- Data was extracted in 5-degree zonal bands, to capture meridional variations. The table below provides some indication of the number of observations for sea surface temperature.
- For each parameter, a histogram was constructed. **In the present article we deal only with sea surface temperature.**

Number of observations of Sea Surface Temperature, 0-50°E

Zonal band	No. of obs West	No. of obs East
10 – 05° N	18378	7769
05 - 00° N	65313	20519
00 – 05° S	70367	66307
05 - 10° S	56913	60142
10 – 15° S	89437	53969
15 – 20° S	160336	172482
20 – 25° S	174186	91688
25 – 30° S	205200	192850
30 – 35° S	267205	191977
35 – 40° S	47441	25868
40 – 45° S	7439	6443
45 – 50° S	2373	2850
50 – 55° S	2238	2037
55 – 60° S	1211	702
60 – 65° S	959	1072
65 – 70° S	1570	9801

Sea surface temperature (SST) characteristics

To consider appropriate QC procedures for sea temperature (SST), it was important to obtain some insight into the variations of this parameter.

Some examples of the SST percentage-histograms are shown in the accompanying figures. Apart from obtaining insight into anomalies in the data, it was found that some useful results were obtained from the process.

- The number of observations vary from north to south, dropping rapidly south of the continent. The increase in observations south of 65°S may be ascribed to (research?) vessels operating in that area, and may be confined to summer.
- In the equatorial region, sea surface temperatures off the east coast are generally lower than off the west coast. This was rather unexpected.
- Further south of the equator, the western sector started becoming cooler than the eastern sector. This could be ascribed to the Benguela Current system and the upwelling off the west coast, in contrast to the western boundary currents off the east coast. This east-west temperature difference reached a maximum between 20° and 30° South, to a magnitude of about 7° C.

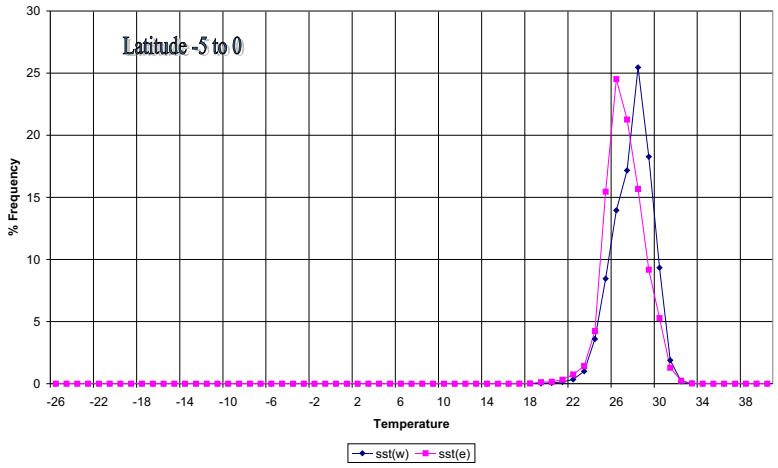
Specific anomalies

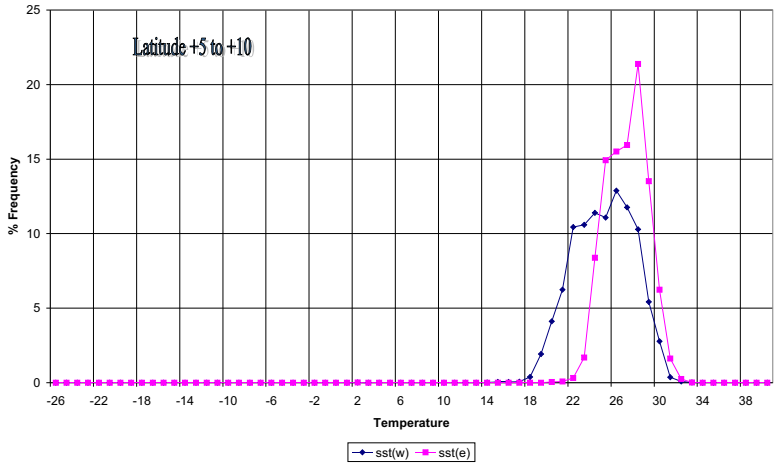
- Not visible in the graphs because of the temperature scale, were some scattered observations located significantly beyond the tentative edges of the peaks. The possibility exists that the "below normal" temperatures that were negative, might have been measured as positive values, but became negative during the encoding and transmission.
- "Below normal" temperatures were much more prominent than the "above normal" values. E.g., while there were anomalous temperatures above 30 degrees (but none above 40), "below normal" values extended all the way to -40 degrees.
- At the same time, the further south the area and the lower the mean temperature, the more prominent became the "above normal" temperatures.

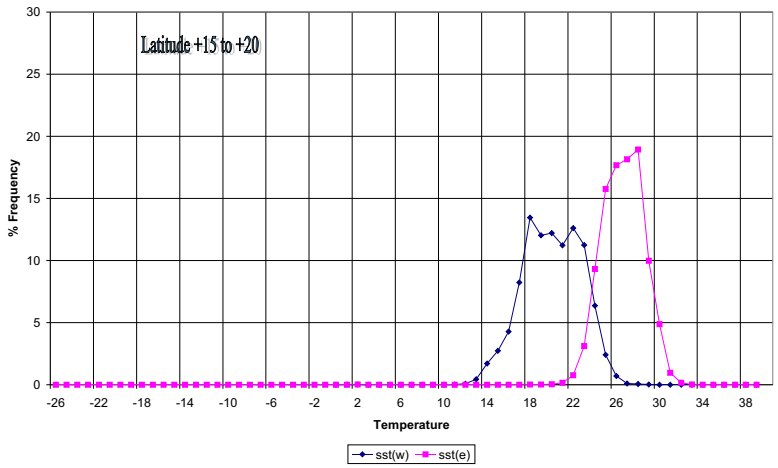
Comments

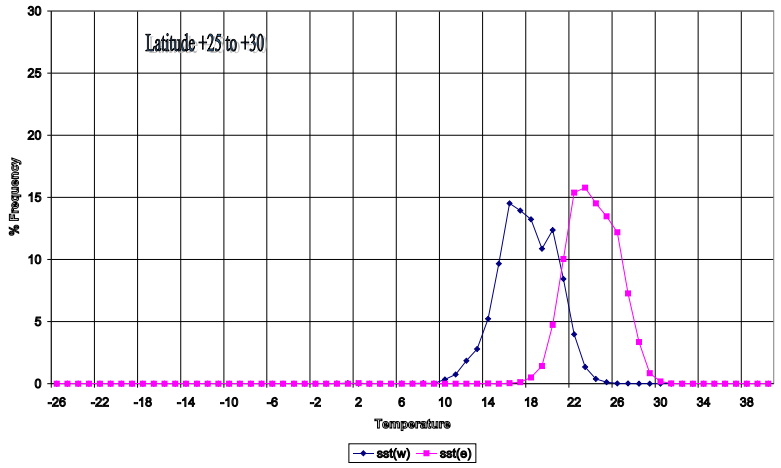
- There are many things that can contribute to anomalous data. A simple shift in the character format, during transmission or reprocessing, could cause 21° to become 2.1°. In addition, the loss of a character in the data stream could cause other parameter values to "shift up" one space.
- Because the mean temperatures do not provide insight into a possible screening mechanism, the behaviour of percentiles was investigated. As an example, the 5% percentiles (2.5% on either side of the peak) were derived (see Figure on page 8). These indicate the lower temperatures along the west coast, due to the Benguela and upwelling. However, there are anomalously high temperatures south of 45° in the "western" sector (but not in the eastern sector). In this area the number of observations are not very high (see Table), so that a limited number of erroneous reports can impact on the validity of the data set as a whole. Inspection of individual readings showed that a single vessel submitted a number of reports in about 55°S over a period of two weeks that were all high (17°C). The highest reported temperature in these latitudes was 30°C!. These values would still be acceptable in terms of the WMO screen (up to 37°C would still be valid).

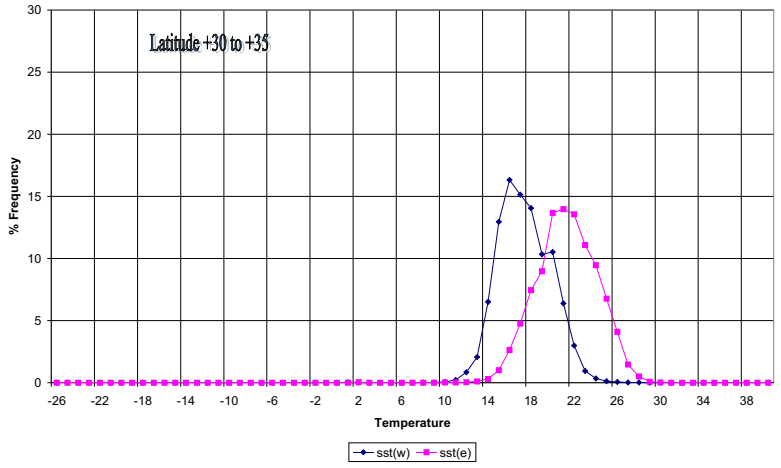
*Percentage histograms for selected zonal bands. Each graph shows the **east** and **west** distribution. Notice the separation of the histograms in 15° to 30°S due to cold-water effects along the west coast, and rejoining south of the continent.*

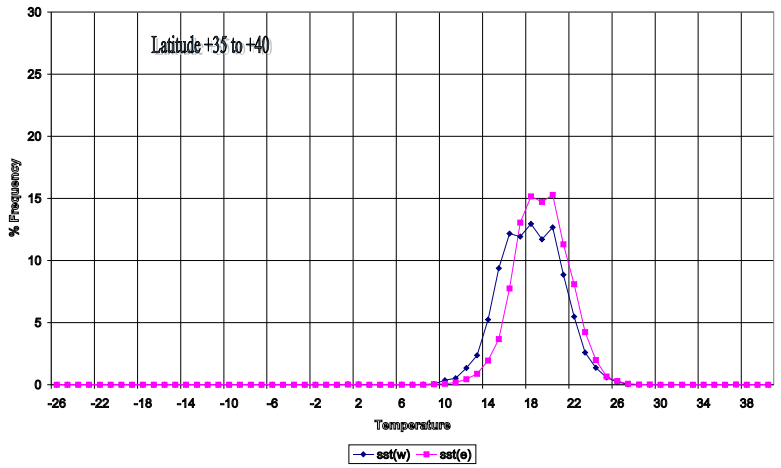







Latitude +25 to +30
% Frequency
Temperature
sst(w)
sst(e)


Latitude +30 to +35
% Frequency
Temperature
sst(w)
sst(e)


Latitude +35 to +40
% Frequency
Temperature
sst(w)
sst(e)

Some aspects of quality control (QC)

Somebody once said: “I don't know how to describe quality, but I'll know it when I see it”

The Oxford dictionary defines “quality” as “degree of excellence”. By “quality control” we therefore mean the *process of defining, improving and ensuring the excellence of something*.

This sounds very simple to achieve, but to show the difficulty, please define (for yourself) the relative quality of these two values: 7.3 and 21.

Yes, it is impossible to determine, because we do not know the units, the measuring instrument, location, time, etc etc. In other words, **to determine the quality of a value (measurement) one needs a wide spectrum of supporting and comparative information.** If this is not available, the quality becomes “unknown” and, by implication, suspect.

The quality of the data in a data centre is of the utmost importance. This data provides the basis for scientific and applied investigations, and thus has an impact on the outcome of many activities. Because we often base our future on what we know about the present, the quality of the data has a direct impact on our prediction of the future.

The definition given above for quality control also implies that there are *degrees of excellence,* suggesting that it depends upon the baseline from which one starts. Also, the quality allocated to something by one person will not necessarily be the same as that allocated by another person. One lab may determine the quality of data at a given level (say “high”), while another lab may regard the quality of the same data as being very different (say “low”).

This raises the question whether there is an “absolute” quality?

No. Quality is site specific, domain specific, purpose-specific and time specific (there may even be others).

Site specific: What is considered as “high quality” in one area or country is not necessarily considered high quality in another area or country.

Domain specific: The quality of sea temperature, as rated by biologists may not be the same when evaluated by a geologist.

Purpose specific: A sea temperature for bathing purposes may be very good if it is accurate to one half °C, while the same temperature may not be useful for physical oceanographic research purposes, where temperatures accurate to 1/100 or 1/1000 °C are required.

Time specific: What is considered good quality today, may be considered of very poor quality 10 or 50 years from now.

What value does quality data really have in the future?

Measurements taken many years ago may have been done on the best equipment of the day, and processed with the best algorithms. However, if this data is used for engineering purposes today (such as design of maritime structures), the

accuracy may be vastly inadequate for the present-day purpose. It would also be difficult to merge data of such low quality (according to today's standards) with more recent data of much higher quality, in an attempt to improve the statistical base.

On the other hand, the longer time series may be useful (at the level of the lowest quality) for studies looking at long-term trends (if the latter exceed the variance of the low-quality data).

Implications of QC

Ensuring the quality of data is both a blessing and a curse.

On the one hand, having better data raises the quality of the conclusions based on the data, reduces risks, etc. This is essential where environmental trends need to be identified earlier and with higher accuracy, where impacts have to be determined, structures designed, etc.

On the other hand, good (quality) data is very expensive, because of the cost of collection, calibration of equipment, processing and checking. The continuous drive to the "perfect" set of data, the newest equipment, and the highest-density measurements adds a considerable price tag to the data. On a global scale, the same financial differentiation that exists between richer and poorer countries seems to permeate into the quality of the data that they can produce.

Conclusion

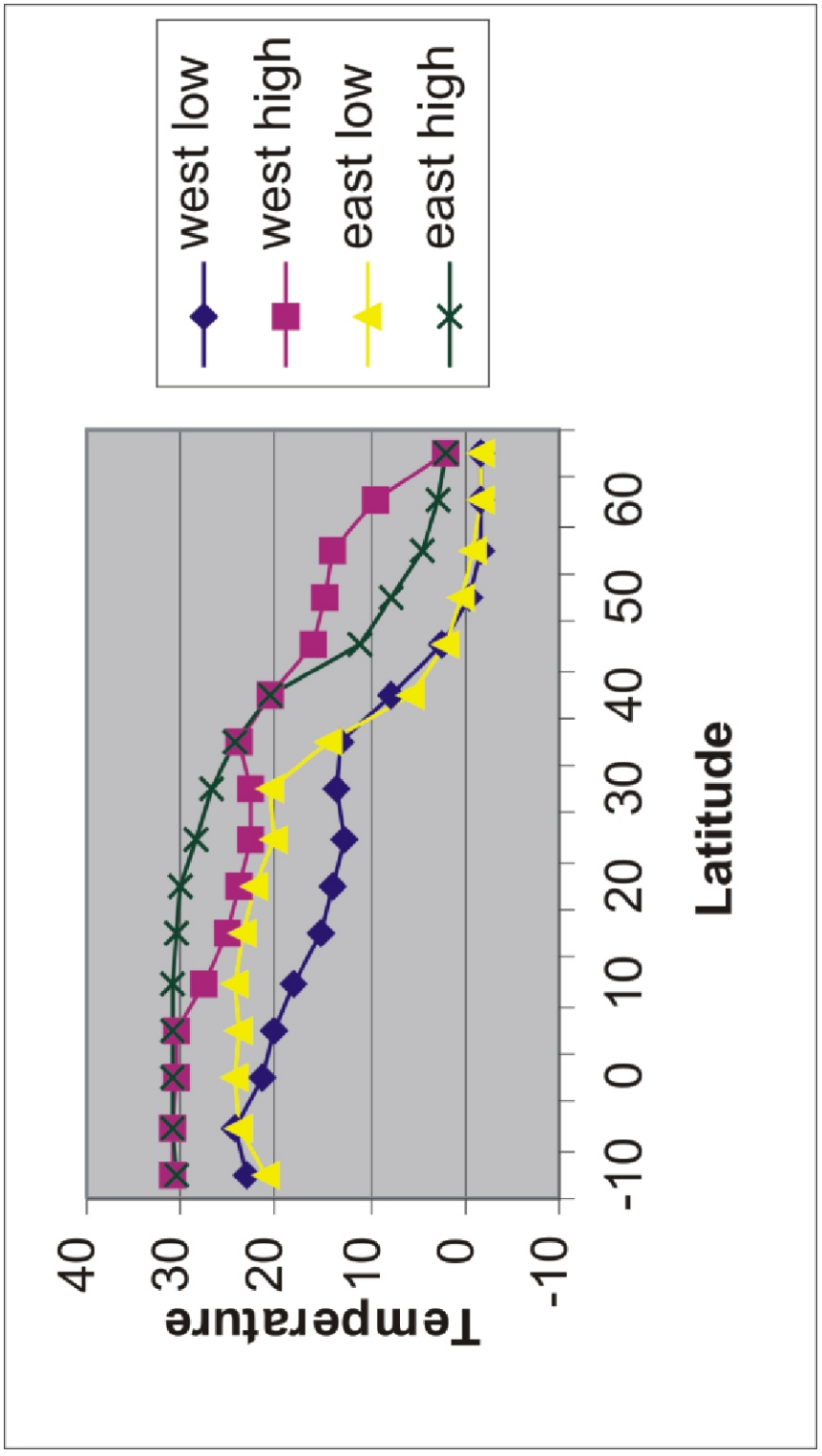
Often, data quality is simply dependant on the attention given to the process of collecting and handling the data, how we look after our equipment, how we ensure that metadata is kept safely, how we ensure that modifications to the data are logged properly.

By ensuring that the quality is the best we can achieve, the "lifetime" of the data is extended, and if this can be achieved in a cost-efficient way, the return on this investment is improved.

Data screens of the World Meteorological Organisation (article page 2)

Data is discarded when lying outside the indicated limits

Parameter	**WMO check**
year	valid year
month	<> 1 - 12
day	no test
hours (GG)	> 23
quadrant (Q)	<> 1,3,5,7
latitude ($L_aL_aL_a$)	abs($L_aL_aL_a$) > 90
longitude ($L_oL_oL_oL_o$)	abs($L_oL_oL_oL_o$) > 180
Overland	no tests done
cloud height (H)	<> 0 - 9
visibility (W)	<> 90 - 99
cloud amount (N)	<> 0 - 9
present weather (ww)	<> various
wind speed indicator (i_w)	<> 0, 1, 3, 4
wind speed (3) (ff)	knots (i_w = 3 or 4) >80, no direction m/s (i_w = 0 or 1) > 40, no direction
wind direction (2) (dd)	> 360, <> 990, no wind speed
air temperature sign (s_n)	<> 0, 1
air temperature (TTT) (drybulb) (2)	< -25, > 40,TTT < dewpoint (DP)
dewpoint sign (s_t)	<> 0, 1, 2, 5, 6, 7 (?)
Dewpoint (DP)	DP > TTT
WetBulb (WB)	WB > TTT
sea surface temp ($T_wT_wT_w$)	< -2, > 37
atmospheric pressure (PPPP)	< 930, > 1050
wave period (P_wP_w)	> 20, <> 99
wave height (½ meters) (H_wH_w)	> 35 (17.5 m)
swell direction ($D_{w1}D_{w1}$)	> 360, <> 990
swell period ($P_{w1}P_{w1}$)	> 25, <> 99
swell height (½ meters) ($H_{w1}H_{w1}$)	> 35 (17.5 m)



Meridional variation of the 5% percentiles (e.g. "west low" is the lower 2.5% percentile of the western segment). Lower values of the western curves, relative to the eastern curves, in the latitudes 0-35°S, are due to the colder water of the Benguela and upwelling off the west coast.