Mediterranean Marine Science
Indexed in WoS (Web of Science, ISI Thomson) and SCOPUS
The journal is available on line at http://www.medit-mar-sc.net

Paper presented at the
"International Conference on Marine Data and Information Systems - IMDIS", Athens, 2008

Managing oceanographic data collated from on-line Information Systems

A. IONA[1], A. THEODOROU[2] and E. BALOPOULOS[1]

[1] Hellenic Centre for Marine Research/Hellenic National Oceanographic Data Centre - HCMR/HNODC, P.O. Box 712, 19013 Anavissos, Hellas
[2] University of Thessaly, Department of Agriculture Ichthyology and Aquatic Environment, Laboratory of Oceanography, Hellas

Corresponding author: sissy@hnodc.hcmr.gr

Abstract

The data management procedures applied to oceanographic data that extracted from major on-line Information Systems are presented. The databases of the World Data Centre (NODC/NOAA), the French Operational Data Centre (Coriolis) and the International Council for the Exploration of the Sea (ICES) were checked for duplicates with the reference MEDAR/MEDATLAS II (2002) database. The retained data standardized to common formats and codes for coherence and compatibility between them as well as with existing datasets already used by the research oceanographic community. The updated dataset of temporal coverage from 1864 to 2007 will be used for the computation of new gridded fields and climatological analysis in the Mediterranean and the Black Seas.

Keywords: Data management; Mediterranean Sea; Quality Assurance; MEDATLAS Format; Climatological Analysis.

Introduction

The most complete existing dataset that has been widely used for statistical analysis and studies of the decadal and the long-term variability of the Mediterranean and the Black Seas is the MEDAR/MEDATLAS II (MEDAR GROUP, 2002) with temporal extension from the begging of the century until 2000. Since 2000, a large amount of additional historical and recent data from research projects and the operational oceanography have been made available at the Web Portals of various International Projects and Data Centres. The main purpose of this work is to obtain a temporally updated dataset since the completion of the MEDAR/MEDATLAS II Project so as to be used for the study of the climatic variability of the Mediterranean and the Black Seas. To accomplish this, vertical profiles of physical and bio-chemical station data extracted from on-line Information Systems, checked for duplicates with the reference MEDAR/MEDATLAS II dataset and reformatted to widespread formats for further processing. The data were organized in files according to their data type. Totally, seven data types were processed: CTD, BOTTLE, Expendable Bathythermographs (XBT/XCTD), Me-

Table 1
Number of stations per data type and data source.

DATA TYPE	DATA SOURCES		
	WOD05	CORIOLIS	ICES
CTD	2743		240
BOTTLE	11817		697
XBT, MBT, XCTD	68515	8291	
Profiling Floats	63	6924	
Drifting Buoys	3167		
Gliders		13789	
Others		361	

chanical Bathythermographs (MBT), Profiling Floats (PFL), Drifting Buoys (DRB) and Gliders data. An additional data type with the characterization "*Others*" consisting of temperature profiles extracted from the Coriolis Data Centre. A quantitative description as well as the geographical distribution of the collated data sets are presented in section 2. The quality assurance procedures that applied to the data sets and some conclusive remarks of the work are described in section 3 and 4 respectively.

Data sets description

The web portals that searched out for additional historical and recent data at the Mediterranean and the Black Seas are available at the following web addresses: a) The World Ocean Database 2005 (WOD05) at http://www.nodc.noaa.gov/cgi-bin/OC5/SELECT/builder.pl, b) The Coriolis Database of the French Operational Data Centre at http://www.coriolis.eu.org/cdc/dataSelection/cdcDataSelections.asp and c) The ICES Oceanographic Database at: http://www.ices.dk/ocean/dotnet/HydChem/HydChem.aspx. A total of 116273 additional vertical profiles of hydrological and bio-chemical stations from 5642 cruises found for the time period 1864-2007. The majority of the data originated from released bathythermographs collected from the navies of various countries before 2000. The source of the Profiling Floats is the international ARGO Program and the source of the WOD05 Drifting Buoys is the Global Temperature-Salinity Profile Program (GTSPP). Gliders data are vertical profiles of temperature and salinity collected by autonomous drifting vehicles coming from MERSEA and MFSTEP Projects. The Coriolis XBT and XCTD data come from research or commercial ships. Table 1 summarizes the number of stations per data type and data source. The geographical representation of the additional datasets is shown at Figure 1.

An inventory of the extracted variables, the GF3 codes (EXTENDED GF3 CODES FOR PARAMETERS AND UNITS, 1987), the full parameter names, their units and the total number of profiles per parameter is shown at Table 2. The geographical and the annual distributions of the core hydrological and bio-chemical parameters are presented at Figure 2.

Quality Assurance

Format Conversions

To ensure coherence and the compatibility between the extracted in different formats data sets , the collated data were reformatted at two ASCII formats depending on the available information that associated with the data: a) the MEDATLAS format (MEDAR-MEDATLAS PROTOCOL, 2001) and b) the Ocean Data View generic spreadsheet format (SCHLITZER, R., 2006). The Coriolis data were extracted at MEDATLAS format without any additional conversions. In absence of delayed mode data, real mode data extracted. The ICES oceanographic data were converted to the ODV format due to the lack of many descriptive information fields

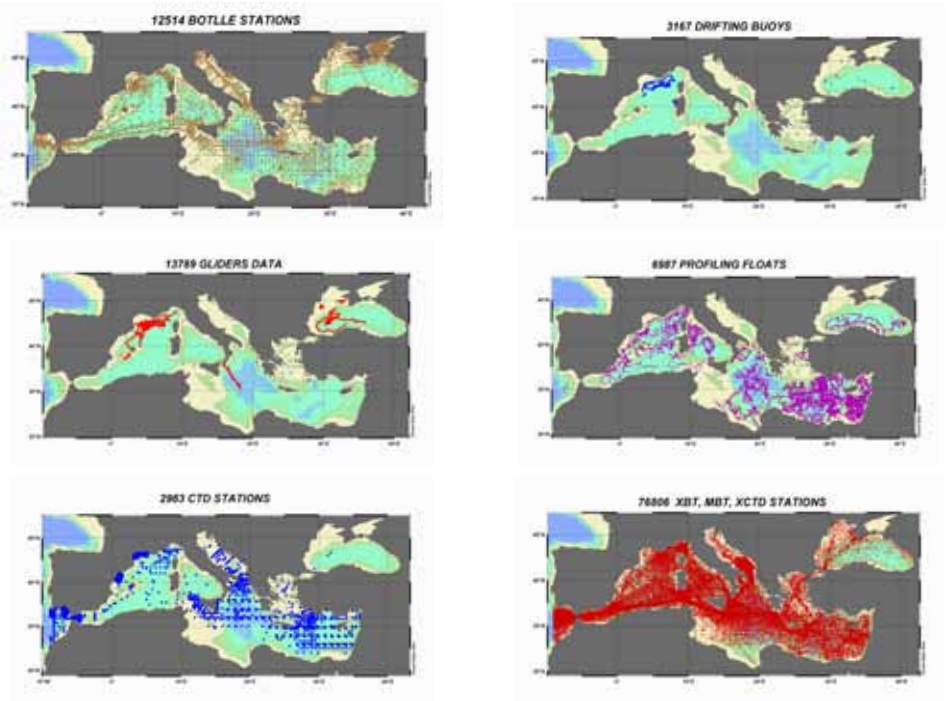

Fig. 1: Geographical distribution per data type.

Table 2
Inventory of the extracted parameters.

GF3 CODE	PARAMETER NAME	UNIT	PROFILES
TEMP	SEA TEMPERATURE	Celsius degree	115715
PSAL	PRACTICAL SALINITY	P.S.U.	29140
DOX1	DISSOLVED OXYGEN	ml/l	4086
PHOS	PHOSPHATE (PO4-P) CONTENT	millimole/m^3	1407
SLCA	SILICATE (SIO4-SI) CONTENT	millimole/m^3	801
NTRA	NITRATE (NO3-N) CONTENT	millimole/m^3	504
NTRI	NITRITE (NO2-N) CONTENT	millimole/m^3	235
AMON	AMMONIUM (NH4-N) CONTENT	millimole/m^3	271
PHPH	PH	pH unit	419
CPHL	CHLOROPHYLL-A TOTAL	milligram/m^3	21
TALK	TOTAL ALKALINITY	Meq/l	60
CHLT	CHLOROPHYLL-TOTAL	milligram/m^3	750
POCP	PARTICULATE ORGANIC CARBON/POC	milligram/m^3	64
CORG	DISSOLVED ORGANIC CARBON	millimole/m^3	41

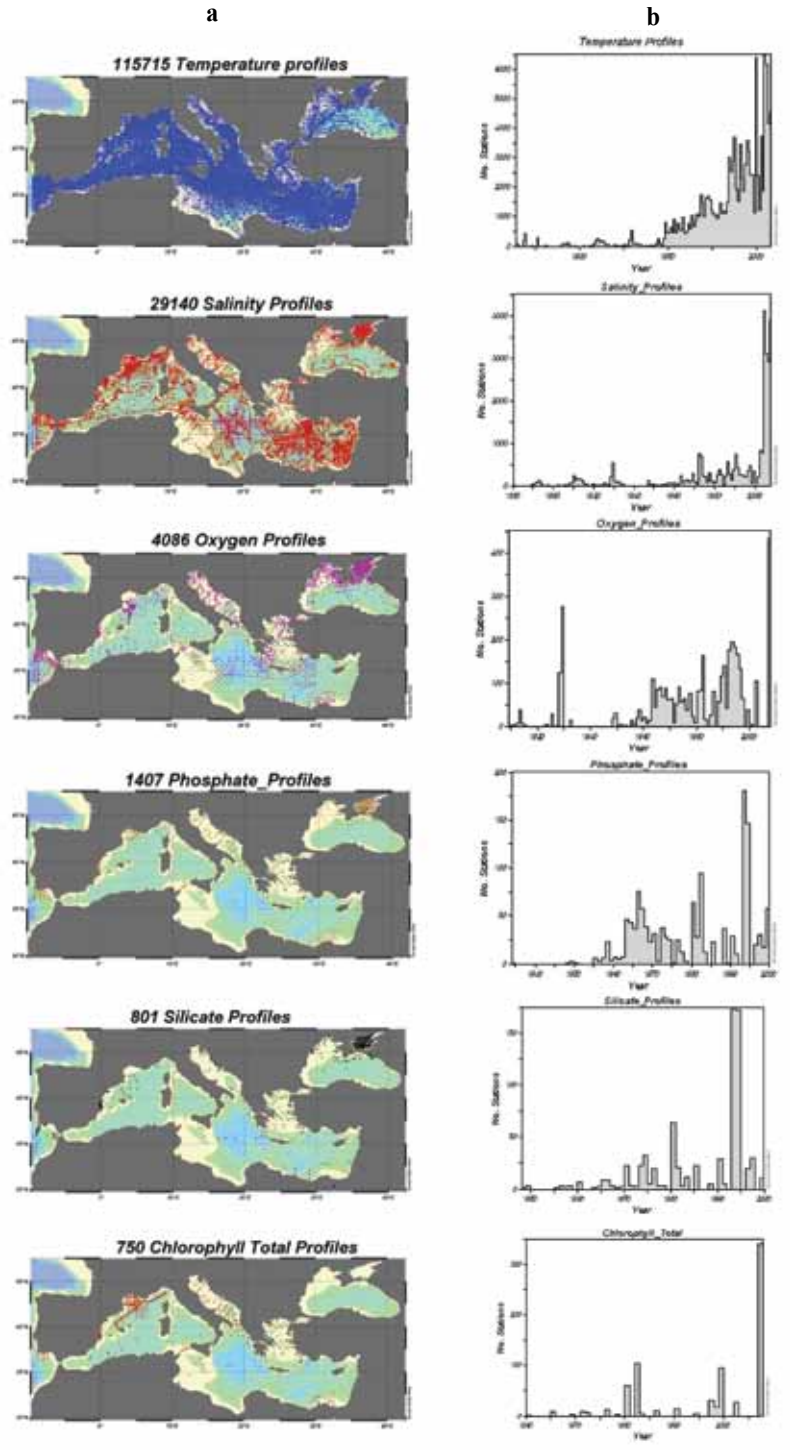

Fig. 2: Distribution of measured parameters a) Geographical and b) annual.

Table 3
Exchange format and number of cruises files.

DATA SOURCES	DATA TYPE	DATA FORMAT	CRUISES FILES
WOD05	CTD	MEDATLAS	72
	BOTTLE	MEDATLAS	797
	XBT, MBT	MEDATLAS	4503
	Profiling Floats	MEDATLAS	17
	Drifting Buoys	MEDATLAS	10
CORIOLIS	Profiling Floats	MEDATLAS	80
	XBT, XCTD	MEDATLAS	85
	Gliders	MEDATLAS	24
	Others	MEDATLAS	6
ICES	CTD	ODV	18
	BOTTLE	ODV	36

that are required in the MEDATLAS format. The WOD05 data were extracted at the native ASCII format (WORLD OCEAN DATABASE, 2005) and then converted into the MEDATLAS. The WOD05 file structure consists of six headers including the following information:

Primary Header: information related to the identification of an individual cast, such as date, time, location, NODC country code, cruise code, and a unique cast number.

Secondary Header: information such as meteorological data, sea floor depth, instrument, ship, institute, and project.

Variable-specific secondary header: information specific to each individual measured variable such as originator's units and methods.

Character Data: Originator's cruise codes, originator's cast codes, and Principal Investigator's code.

Biological Header: information necessary to understand how biological data were sampled. "Biological" data are defined as plankton biomass (weights or volumes) and taxa-specific observations.

Taxa-specific and Biomass Data: Plankton weights, volumes, and/or concentrations, for an entire sample (biomass) or for individual groups of organisms (taxa-specific).

The above information -wherever existed, included to the final data set. The exchange format per data type and data source as well as the number of the cruises files are shown at Table 3.

Checks for duplicates

The duplicates between the different data sources were identified by comparing the position and the date (and time) of the profiles and were eliminated from the final dataset. The choice of the criteria for the duplicates detection was based on the experience of the MEDAR/MEDATLAS II Project. In this work, the following criteria were used:

- Space interval: 1 nautical mile
- Time interval: 1 day

The stations extracted from the three web portals automatically checked with the MEDAR/MEDATLAS II stations and rejected as duplicates if their spatial difference was less than 1 nm and their temporal difference less than 1 day.

Replicates detected between WOD05 and MEDAR/MEDATLAS II with time difference varying from zero until six hours. The use of a smaller temporal threshold value than 1 day (e.g. 15 minutes) maintained a lot of duplicates in the system. The use of a spatial threshold value greater than 1 nm (e.g. 5 nm based on the example of the GTSP Project, eliminated a lot of good data.

Duplicates checks performed between same data types, e.g. WOD05 CTD datasets checked with MEDAR/MEDATLAS II CTD data, otherwise Bottle and/or other data types rejected automatically as doubles with the CTD casts. Cross checking among different data types e.g.

Table 4
Mapping of WOD05 Data Quality Flags to MEDATLAS QC Flag scale.

WOD05 Data Quality Flags		Corresponding MEDATLAS Quality Flag	
Depth Flags			
0	accepted value	**1**	Correct Value
1	duplicates or inversions in recorded depth (same or less than previous depth)	**4**	False Value
2	density inversion	**4**	False Value
Observed Level Flags			
0	accepted value	**1**	Correct Value
1	range outlier (outside of broad range check)	**4**	False Value
2	failed inversion check	**4**	False Value
3	failed gradient check	**3**	Dubious Value
4	observed level "bullseye" flag and zero gradient check	**4**	False Value
5	combined gradient and inversion checks	**4**	False Value
6	failed range and inversion checks	**4**	False Value
7	failed range and gradient checks	**4**	False Value
8	failed range and questionable data checks	**4**	False Value
9	failed range and combined gradient and inversion checks	**4**	False Value

MBT with XBT performed manually only to a limited sample of the data without revealing any duplicates.

In the ICES oceanographic database, there are Bottle data being replicates of low resolution CTD. Thus, the ICES data types checked with all the MEDAR/MEDATLAS II, WOD05 and Coriolis data types.

Quality Flags

In the MEDATLAS format there are included two categories of quality flags that are assigned to the data sets by the data originators or by the Data Centres: the metadata and the data values quality control flags. The WOD05 database does not include metadata quality flags. Thus, their construction was carried out in two stages. Firstly, all the header flags assigned to zero and then, after geographical representation and elimination of the on land located stations, the latitude and longitude quality control values assigned to 1 (correct value). Since the Coriolis datasets were extracted in MEDATLAS format no further reformatting was performed. The ODV format does not include metadata quality flags and no further processing was applied.

The mapping between the World Ocean Database Quality Flags and of the MEDATLAS Flag Scale is shown in Table 4.

The WOD05 quality flags for the entire cast as a function of parameter have not been maintained. Instead, the global parameter quality flag is assigned as follows:

- 20 % of the flagged values for high resolution data (e.g. CTD casts)
- 50 % of the flagged values for low resolution data (e.g. Bottle casts)

For the ICES data, two values are used, 1 (Unknown) and 8 (as missing value) according to the ODV QC flag scale.

Conclusions

More than 110.000 additional of the MEDAR/MEDATLAS II database vertical profiles of physical and bio-chemical parameters were extracted from three public domain databases and standardized to internationally agreed standards and formats in order to obtain an updated merged dataset to perform climatological analysis in the Mediterranean and the Black Seas. Special concern was given to the translation of the quality flags from one format to another because only "good" data should be used to statistical computations. The data are being gradually loaded to the database of Hellenic National Oceanographic Data Centre and are available on-line in MEDATLAS format (http://hnodc.hcmr.gr/services.html).

Acknowledgements

The work has been carried out within the framework of the EU/SESAME Integrated Project (Contract No. GOCE-2006-036949) and the EU/CIRCE Integrated Project (Contract No. 036961). The collated data will be used within the EU/FP6 SEADATANET Integrated Infrastructure Initiative Contract No. 026212) for the computations of the Mediterranean Sea climatologies. We acknowledge the financial support of the European Commission's Sixth Framework Programme. We thank Dr Alexey Mishonov from NODC/NOAA, USA and Catherine Maillard and Michele Fichaut from IFREMER/SISMER, FRANCE for their valuable comments and support.

References

EXTENDED GF3 CODES FOR PARAMETERS AND UNITS, UNESCO/IOC/IODE, 1987. GF3 A General Formatting System for Geo-Referenced Data - Manual and Guides 17.

MEDAR GROUP, 2002. Mediterranean and Black Sea Database of Temperature, Salinity and Biochemical Parameters and Climatological Atlas [4 CD-ROMs], Ifremer Ed., Plouzane, France, (Available at http://www.ifremer.fr/medar).

MEDAR-MEDATLAS PROTOCOL, 2001 Part I: Exchange format and quality checks for observed profiles", V3, http://www.ifremer.fr/medar/qc_doc/medman_v3.doc.

SCHLITZER, R., 2006. Ocean Data View, http://odv.awi.de.

WORLD OCEAN DATABASE DOCUMENTATION, 2005. Ed. Sydney Levitus. NODC Internal Report 18, U.S. Government Printing Office, Washington, D.C., 163 pp.