



SeaDataNet

PAN-EUROPEAN INFRASTRUCTURE FOR
OCEAN & MARINE DATA MANAGEMENT

FIRST RELEASE OF THE AGGREGATED DATA SETS PRODUCTS

WP10 SECOND YEAR REPORT - DELIVERABLE D10.2

Project Acronym : SeaDataNet II

Project Full Title : SeaDataNet II: Pan-European infrastructure for ocean and marine data management

Grant Agreement Number : 283607



First release of the aggregated data sets products – Friday 21 July 2017
sdn-userdesk@seadatanet.org – www.seadatanet.org

Deliverable number	Short Title
D10.2	Aggregated data sets
Long title	
First Release of the Aggregated Data Sets Products	
Short description	
<p>This document describes the procedure implemented for the generation of the SeaDataNet 2 versions of the aggregated products. Detailed descriptions of the data sets per sea region (Mediterranean Sea, Black Sea, Arctic Sea, Baltic Sea, North Sea and the North Atlantic Ocean) are given with the general description of the dataset, the data quality assessment procedure and results.</p> <p>Aggregated datasets are available under ftp://ftp.ifremer.fr/ifremer/sismer/SeaDataNet2/Products/</p>	
Author	Working group
S. Simoncelli, C. Coatanoan, V. Myroshnychenko, H. Sagen, Ö. Bäck, S. Scory, M. Tonani, A. Grandi, R. Schlitzer, M. Fichaut	WP10

History

Version	Authors	Date	Comments
1.0	S. Simoncelli, M. Tonani, A. Grandi,	10/12/2013	Creation
1.1	C. Coatanoan, V. Myroshnychenko, H. Sagen, Ö. Bäck, S. Scory	09/22/2014	Regional reports
1.2	C. Coatanoan, V. Myroshnychenko, H. Sagen, Ö. Bäck, S. Scory	10/20/2014	Revised regional reports
1.3	M. Fichaut		comments
1.4	S. Simoncelli	10/31/2014	Corrections, harmonization
1.5	M. Fichaut	04/11/2014	Finalisation of the document
1.6	M. Fichaut	12/11/2014	Add Black sea tables (Tab.6 and Tab.7) Add link to ftp product directory

Table of contents

1. Introduction.....	11
1.1. Objectives of the second year of activity	11
1.2. Time Schedule and Quality Control Strategy	11
1.3. SDN-MyOcean Collaboration	12
2. Data Aggregation Procedure	14
3. Quality Assessment Process	15
4. WP10 Quality Control.....	18
5. MyOcean INSTAC Quality Control and feedback	19
6. SDN analysis of INSTAC Anomalies.....	22
7. Results of the first aggregation and quality assessment	22
8. Second data aggregation and quality assessment analysis	23
9. Regional Aggregated Data Sets	24
9.1. Mediterranean Sea.....	24
9.1.1. General Characteristics of the Mediterranean Sea historical data set.....	24
9.1.2. Atlantic Box	29
9.1.3. Marmara Sea	31
9.1.4. The Mediterranean Sea restricted data collection	34
9.2. Black Sea.....	35
9.2.1. General Characteristics of the Black Sea historical data set.....	35
9.2.2. Data Quality assessment procedure for the Black Sea	40
9.2.3. Quality assessment results.....	40
9.2.4. The Black Sea restricted data collection	41
9.2.5. Conclusions.....	42
9.3. Arctic Sea	44
9.4. Baltic Sea	46
9.4.1. General Characteristics of the Baltic Sea historical data set	46
9.4.2. Data Quality assessment procedure for The Baltic Sea	49
9.4.3. Conclusions.....	49
9.5. North Sea.....	51
9.5.1. General characteristics of the North Sea Historical data collection	51
9.5.2. Quality of the historical dataset.....	55
9.5.3. Quality of the restricted dataset.....	56

9.5.4. Conclusions.....	57
9.6. The North Atlantic Ocean.....	58
9.6.1. General Characteristics of the North Atlantic Ocean historical data set.....	58
9.6.2. The North Atlantic Restricted dataset.....	61
9.6.3. Conclusions.....	62
10. General Conclusions on the quality of the aggregated data sets	63
References.....	66

List of Figures

Figure 1 Schema of the Quality Check strategy implemented by WP10 in synergy with NODCs and MyOcean INSTAC.....	12
Figure 2 Time schedule and list of actions during the second year of WP10 activities.....	13
Figure 3 Review of WP10 actions according the decision to postpone the release of V1.1 historical data collection.	13
Figure 4 Results of the duplicate check: number of duplicates and actions to be undertaken by data providers.....	14
Figure 5 Timeline of the feedback procedure between MyOcean INSTAC quality assessment results on TS data collections and SDN CDI partners between June and September 2013.	17
Figure 6 Table of the list of anomalies where CDI partners had to insert their comments. ...	17
Figure 7 Example of table of the list of updated CDI asked from CDI partners after data anomalies check procedure.	18
Figure 8 Template where CDI partners have to summarize the QC action, the anomalies detected by MyOcean INSTAC and the number of resulting true anomalies.....	18
Figure 9 Building V3 and V4 MyOcean reprocessed T&S in situ products.....	20
Figure 10 Schematic of the new branch history on MyOcean INSTAC ftp, which contains SDN data.....	20
Figure 11 Screenshot from MyOcean Interactive Catalogue describing the Mediterranean product and its data access.....	21
Figure 12 Outline of MyOcean type of anomalies as analyzed by WP10 (<i>by C.Coatanoan</i>). ..	22
Figure 13 Temperature and Salinity data collection for the Mediterranean Sea in the time period 1900-2012: (a) Data distribution map; (b) Data density map.	25
Figure 14 (a) Annual data distribution and (b) seasonal data distribution for the time period 1900-2012 in the Mediterranean Sea.	26
Figure 15 Data distribution map for the Mediterranean Sea: (a) Temperature, (b) Salinity..	26
Figure 16 Temperature scatter plots: (a) Data with QF1; (b) data with QF0.....	28
Figure 17 Salinity scatter plots: (a) data with QF1; (b) data with QF0.....	28
Figure 18 Mediterranean TS data collection for the time period 1900-2012: (a) TS diagram of full data collection; (b) TS diagram considering only data with QF1 (good) and QF2 (probably good); (c) TS diagram considering only data with QF0 (no quality control)....	29
Figure 19 Zooms from the Mediterranean Sea data distribution map that unveil the presence of observations with erroneous locations on land.	29
Figure 20 Data distribution map (a) and data density map for the Atlantic box included in the Mediterranean historical data collection. (31322 stations)	30
Figure 21 Annual (a) and seasonal (b) data distribution in the Atlantic box.	30
Figure 22 Temperature scatter plots for the Atlantic box: (a) entire data set; (b) data with QF=1; (c) data with QF=0.....	31
Figure 23 Salinity scatter plots for the Atlantic box: (a) entire data set; (b) data with QF=1; (c) data with QF=0.	31
Figure 24 (a) Data distribution map and (b) data density map for the Marmara Sea.	32
Figure 25 Annual (a) and seasonal (b) data distribution in the Marmara Sea.	32

Figure 26 Temperature scatter plots for the Marmara Sea data set: (a) entire data set; (b) data with QF=1.	33
Figure 27 Salinity scatter plots for the Marmara Sea data set: (a) entire data set; (b) data with QF=1.	33
Figure 28 The Mediterranean Sea restricted data collection: (a) Data distribution map; (b) Data density map; (c) annual distribution map.	34
Figure 29 Data distribution map: Temperature at the surface (98870 stations).....	35
Figure 30 Data distribution map: Salinity at the surface (93688 stations)	36
Figure 31 Annual distribution of stations with Temperature observations (left column represents all stations before 1958)	36
Figure 32 Monthly distribution of stations with Temperature observations	37
Figure 33 Seasonal distribution of stations with Temperature observations.....	37
Figure 34 Vertical distribution of Temperature data along the IODE standard levels	38
Figure 35 Temperature and Salinity scatter plots for entire dataset before QC	38
Figure 36 Typical Temperature profiles	39
Figure 37 T-S diagram for entire dataset: a) scatter plot; b) typical shape of T-S curve	39
Figure 38 The Black Sea restricted data collection: (a) Data distribution map; (b) Data density map; (c) Annual distribution map.	42
Figure 39 Maps of data coverage for different IODE depth levels: a) 20m; b) 50m; c) 100m; d) 500m; e) 1000m; f) 2000m.....	43
Figure 40 Arctic1900-2013 TS data distribution map V1.1 aggregation for the time period 1900-2013	44
Figure 41 Arctic1900-2013 TS data density map V1.1 aggregation for the time period 1900-2013.....	45
Figure 42 Arctic1900-2013 (a) Annual data distribution and (b) seasonal data distribution for the time period 1900-2013	45
Figure 43 Arctic1900-2013 TS data collection for the time period 1900-2013: (a) TS diagram after range check analysis; (b) TS diagram considering only data with QC flags = 1 (good) and 2 (probably good) for depth, T and S; (c) TS diagram considering only data with QC flags = 0 (no quality control) for depth, T and S.....	45
Figure 44 TS data collection for the Baltic Sea in the time period 1990-2013: (a) Data distribution map; (b) Data density map.....	46
Figure 45 (a) Annual data distribution and (b) seasonal data distribution for the time period 1990-2013 in The Baltic Sea.	47
Figure 46 The Baltic Sea restricted data collection. (a) Data distribution map, (b) data density map and (c) annual distribution map.....	48
Figure 47 Baltic Sea TS data collection for the time period 1990-2013: (a) TS diagram after QC; (b) TS diagram before QC.	48
Figure 48 (a) Salinity variation in The Baltic Sea, (b) sub-regions used to easier handle the quality control.	49
Figure 49 Location of stations in the TS data collections for the North Sea for the period 1900-2013: (a) freely accessible data (“historical dataset”), (b) restricted dataset.....	51
Figure 50 Data density maps of the TS data collections for the North Sea for the period 1900-2013: (a) historical dataset (b) restricted dataset.	52

Figure 51 Time histogram of the data in the historical dataset: (a) full dataset, (b) focus on 1988 (200343 CDIs), (c) focus on 1989 (317290 CDIs). 52

Figure 52 Data density maps of the historical dataset: (a) focus on 1988 (200343 data points), (b) focus on 1989 (317290 data points). 53

Figure 53(a) Data density maps of the historical dataset, from 1900 till 2013, excluding 1988 and 1989; (b) distribution over time of the data in the historical dataset, from 1900 till 2013, excluding 1988 and 1989. 53

Figure 54 (a) Location of the stations in the restricted dataset for the period 1989–1990, (b) Distribution of data over time in the restricted dataset for the period 1989–1990. 54

Figure 55 (a) Location of the stations in the restricted dataset for the period 1993–1995, (b) Distribution of data over time in the restricted dataset for the period 1993–1995. 54

Figure 56 Seasonal distribution of the data: (a) historical dataset, (b) restricted dataset..... 55

Figure 57 North Sea TS historical data collection: (a) TS diagram after range check analysis; (b) TS diagram considering only data with QC flags = 1 (good) and 2 (probably good) for T and S; (c) TS diagram considering only data with QC flags = 0 (no quality control) for T and S. 55

Figure 58 Temperature data from the North Sea historical data collection where original QF=1, exhibiting data that obviously weren't correctly flagged..... 56

Figure 59 North Sea TS restricted data collection: (a) TS diagram of the full dataset; (b) TS diagram considering only data with QC flags = 1 (good) and 2 (probably good) for T and S; (c) TS diagram considering only data with QC flags = 0 (no quality control) for T and S. 57

Figure 60 TS data collection for The North Atlantic Ocean in the time period 1900-2013: (a) Data distribution map; (b) Data density map. 58

Figure 61 (a) Annual data distribution and (b) seasonal data distribution for the time period 1900-2013 in The North Atlantic Ocean. 58

Figure 62 North Atlantic TS data collection for the time period 1900-1999: (a) TS diagram after range check analysis; (b) TS diagram considering only data with QC flags = 1 (good) and 2 (probably good) for T and S; (c) TS diagram considering only data with QC flags = 0 (no quality control) for T and S. 59

Figure 63 North Atlantic TS data collection for the time period 2000-2008: (a) TS diagram after range check analysis; (b) TS diagram considering only data with QC flags = 1 (good) and 2 (probably good) for T and S; (c) TS diagram considering only data with QC flags = 0 (no quality control) for T and S. 59

Figure 64 North Atlantic TS data collection for the time period 2009-2013: (a) TS diagram after range check analysis; (b) TS diagram considering only data with QC flags = 1 (good) and 2 (probably good) for T and S; (c) TS diagram considering only data with QC flags = 0 (no quality control) for T and S. 60

Figure 65 Restricted TS data collection for the North Atlantic Ocean in the time period 1900-2013: (a) Data distribution map; (b) Annual data distribution. 62

Figure 66 Restricted TS data collection for The North Atlantic Ocean in the time period 1900-2013: (a) TS diagram considering only data with QC flags = 1 (good) for T and S; (c) TS diagram considering only data with QC flags = 0 (no quality control) for T and S. 62

Figure 67 Example of the new data coverage by Marine Institute (Ireland) between V1 2013 (left) and V1.1 2014 (right). 64

Figure 68 Quality Check procedure that is required before a new data aggregation procedure in order to provide the best SDN V2 final data collections..... 65

List of Tables

Tab. 1	Number of data points for Temperature, Salinity and TS couples For the Mediterranean Sea only.....	27
Tab. 2	Number of data points for Temperature and Salinity and their subdivision according to Quality Flags (QF) 0, 1, 2 and from 3 to 9.	27
Tab. 3	Number of data points for Temperature and Salinity data within the Atlantic box in Figure 20 and their subdivision according to Quality Flags (QF) 0, 1, 2 and from 3 to 9.	31
Tab. 4	Number of data points for Temperature and Salinity data within the Marmara Sea in Figure 24 and their subdivision according to Quality Flags (QF) 0, 1, 2 and from 3 to 9.	33
Tab. 5	Number of Temperature and Salinity data points for the Mediterranean Sea restricted data collection and their subdivision according to Quality Flags (QF) 0, 1, 2 and from 3 to 9.....	34
Tab. 6	Number of stations in the public Black sea data set.	35
Tab. 7	Number of stations in the restricted Black sea data set.	41
Tab. 8	Number of Temperature and Salinity data points for the North Sea historical data collection and their subdivision according to Quality Flags (QF) 0, 1, 2 and from 3 to 9.	56
Tab. 9	Number of Temperature and Salinity data points for the North Sea restricted data collection and their subdivision according to Quality Flags (QF) 0, 1, 2 and from 3 to 9.	57
Tab. 10	North Atlantic TS measurements collection for the time period 1900-2013: number of measurements (and percent) for temperature and salinity sorted by quality flag.....	60
Tab. 11	Number of data for V1 regional data collections, V1.1 data collections and the relative percentage of data increase.	63
Tab. 12	List of EDMO codes of data centers that mostly provided data between V1 and V1.1 in the North Atlantic region.	63
Tab. 13	Number of stations in V1.1 regional data collections which contain only free access data and number of stations in the restricted data collections extracted from SDN data base during the second aggregation phase. The right column presents the percentage of the restricted data versus the total data.	64

1. Introduction

1.1. Objectives of the second year of activity

Main objectives of the second year of WP10 activity were:

1. Aggregation of regional data sets (unrestricted data or under SDN license) for the selected data types (temperature and salinity);
2. To analyze the density and resolution (space, time, depth) of the regional aggregated data sets;
3. To assess the quality of the regional data sets making use of the ODV tool;
4. To report to the relevant SDN national nodes on the quality and possible shortcomings of the data sets and iterate when necessary;
5. To deliver regional aggregated data sets to specific user communities, MyOcean In situ TAC, as agreed;
6. To release the regional aggregated data sets.

1.2. Time Schedule and Quality Control Strategy

A Quality Control strategy, schematized in Figure 1, has been developed by WP10 Regional Coordinators in synergy with MyOcean In-Situ Thematic Assemble Centre (INS-TAC) in order to improve the SeaDataNet data base content and to create the best product deriving from SDN data. The QC strategy involves the NODCs that on the base of WP10 data quality assessment outcome and MyO feedback have to check and eventually correct the original data. The QC procedure has been designed to be iterative in order to facilitate the update and improvement of SDN database content.

Figure 2 and Figure 3 show the time schedule of WP10 activities related to data aggregation and quality assessment.

WP10 activities came along in communication with the ongoing project MyOcean2 in order to develop a true synergy at regional level and create the best historical data sets and serve operational oceanography and climate change communities.

WP10 leader presented these activities at IMDIS Conference (*Simoncelli et al. 2013*). Data aggregation and quality control procedures will be described in the next paragraphs.

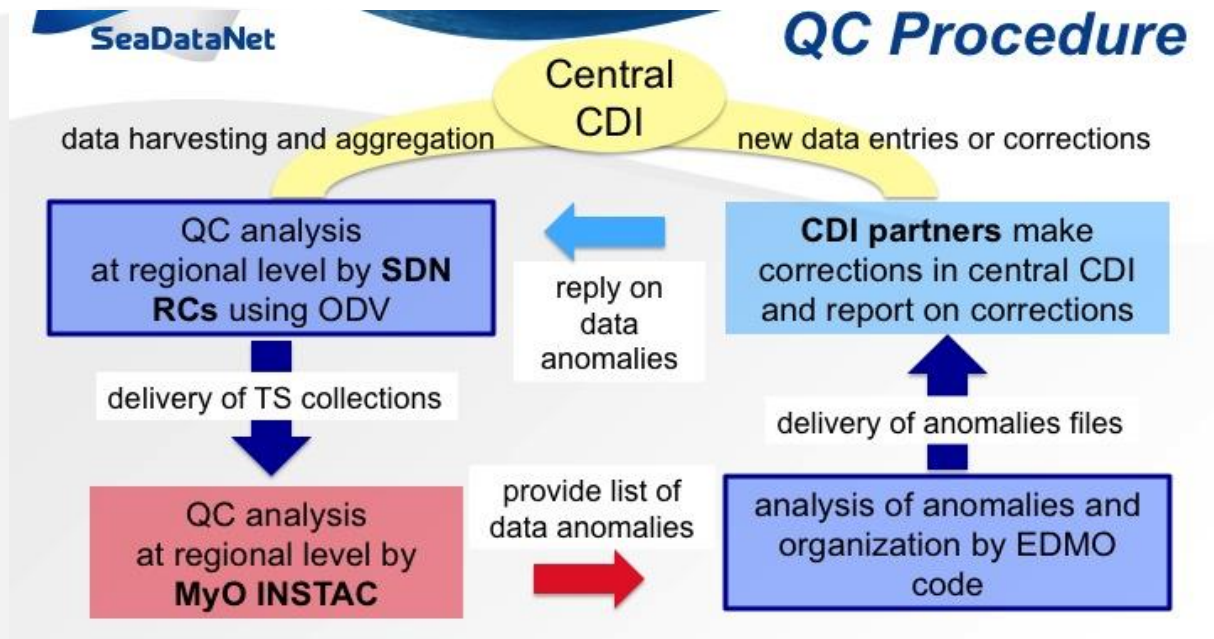


Figure 1 Schema of the Quality Check strategy implemented by WP10 in synergy with NODCs and MyOcean INSTAC.

1.3. SDN-MyOcean Collaboration

The objective of SDN-MyOcean collaboration, stated in a Memorandum of Understanding, is to create TS historical data collections in synergy with MyOcean In-Situ Thematic Assemble Centre (INSTAC) to support and promote monitoring, modelling and downstream service development. In order to generate jointly regional TS products for 20 years (1990-2010) we had to define the interactions between the two Projects. This was done during the 1st Joint Meeting held in Rhode on September 18th 2012, during which we defined:

- common time schedule;
- data flows;
- information exchanges;
- QC strategies;
- Interfaces.

In April 2013 a 2nd Joint Meeting took place in Cork, where SDN RCs and INSTAC regional responsible met to discuss the first outcome on the quality of the first aggregated data sets.

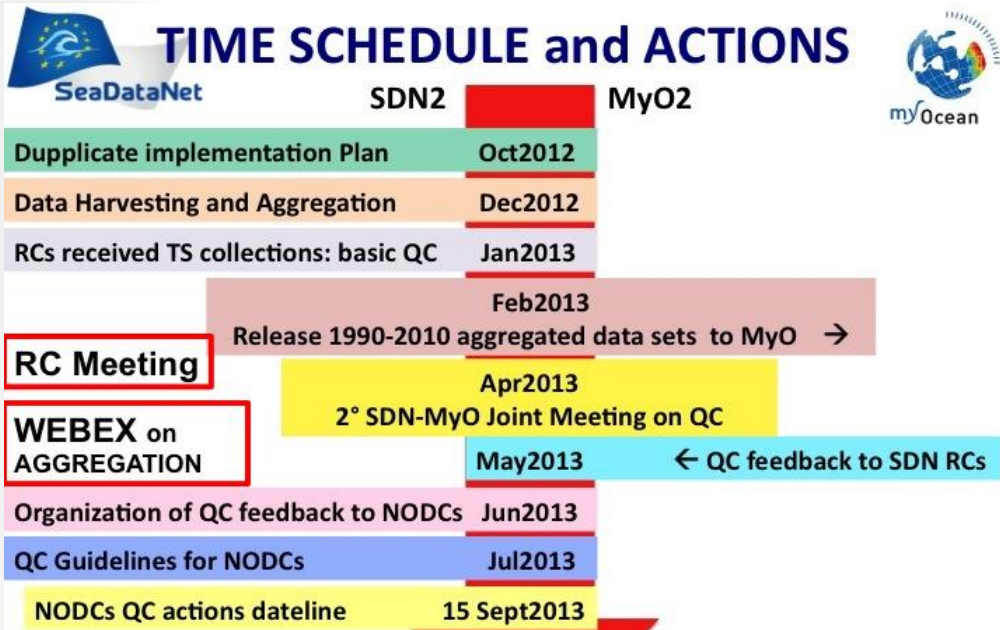


Figure 2 Time schedule and list of actions during the second year of WP10 activities

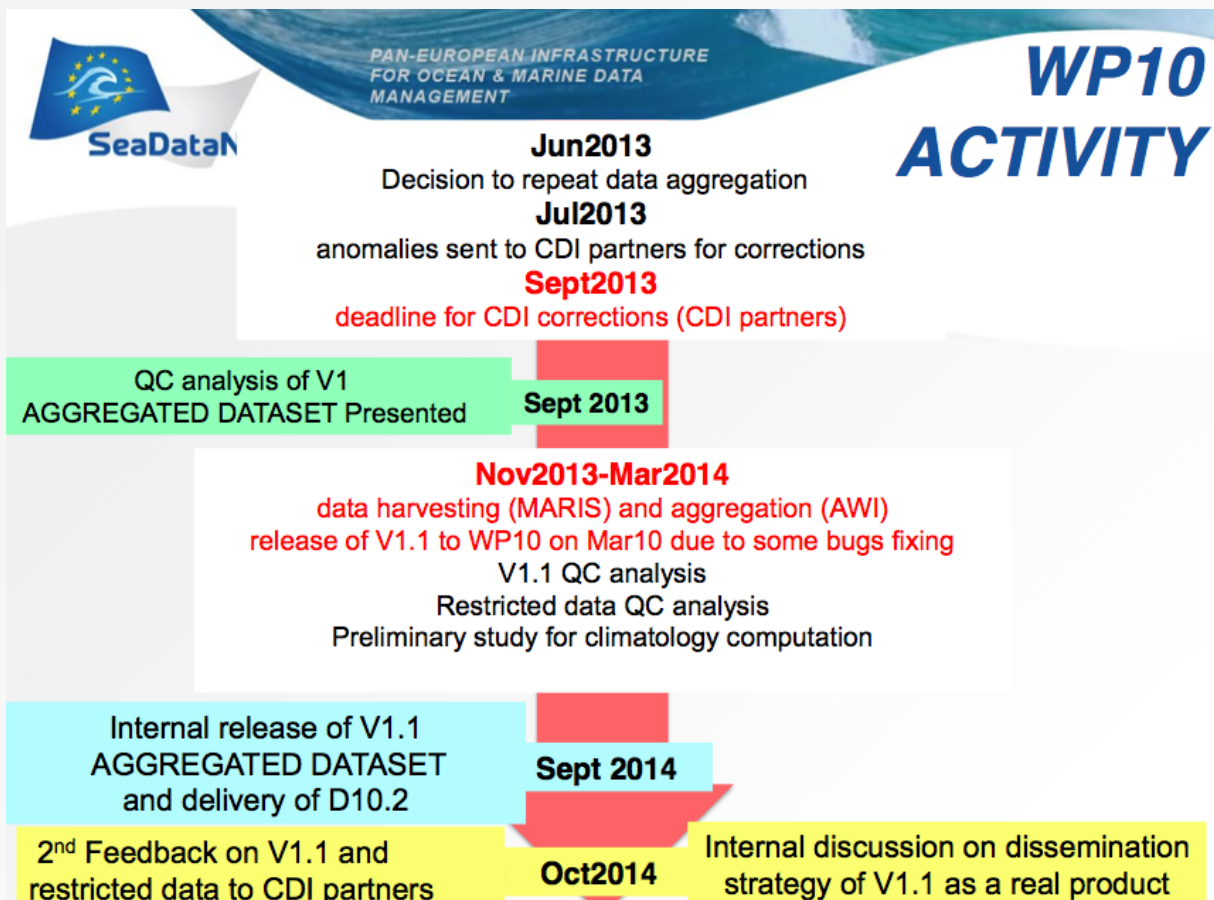


Figure 3 Review of WP10 actions according the decision to postpone the release of V1.1 historical data collection.

2. Data Aggregation Procedure

Data aggregation started in October 2012, after the First Annual Meeting, and needed the coordination between different WPs and Institutes:

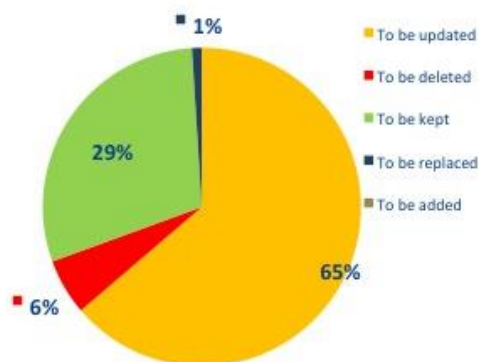
- **WP4** duplicate implementation plan (HCMR);
- **WP5** CDI robot harvesting (MARIS);
- **WP9** aggregation into a single TS Data Collection using SDN Importer of ODV 4.5.3 (AWI);
- **WP10** quality assessment by RCs in collaboration with MyO INSTAC;
- **WP7** network of National Oceanographic Data Centres (NODC's) to correct data anomalies;

First phase started with the **Duplicates Implementation Plan** (WP4 - Sissy Iona, HCMR) based on ODV duplicates checks. It was asked data providers:

- To identify duplicates
- To clean the data (delete, update, replace...)
- To explain their actions in details

After evaluation of all data modifications both the CDI central catalogue and local archives were updated and guidelines were sent to all partners to avoid similar cases in the future. The idea was to create a white list of the cleaned and checked CDIs in order to verify new entries in CDI central catalogue against this list and avoid future duplicates. Figure 2 shows the result of the duplicate check. Only 6% of the potential duplicates were real ones, while 65% needed corrections and 29% were not duplicates.

21 Partners	Potential duplicates	To be updated	To be deleted	To be kept	To be replaced	To be added
Total	60866	38793	3475	17989	596	13



Potential duplicates

- 6% were real duplicates
- 65% needing correction
- 29% not duplicates

Figure 4 Results of the duplicate check: number of duplicates and actions to be undertaken by data providers.

Second phase, **Data Harvesting** of temperature and salinity files by CDI Robot, started in December 2012. *MARIS (WP5 - Dick Schaap)* developed a robot user that uses the CDI Data Discovery and Access Service to query, shop and retrieve data sets from the SDN distributed data centres in automatic way. Data query consisted in searching for all data sets with T&S, whose access was unrestricted or under SeaDataNet License (**~860.000 CDIs**). The Robot was triggered to start harvesting the related ODV files from the distributed data centres through the general CDI shopping mechanism (RSM–DM).

This procedure was essential to test/tune the performance of the RSM (Request Status Manager) - DM (Download Manager) process and find the optimum data requests management procedure. The RSM is fault proof and keeps track of all data requests repeating them in case of disturbances at DM level. In January 2013 a DVD with all the ODV files was delivered to AWI in a storage structure with the full CDI metadata as CSV file.

In the third phase AWI (*WP9 - R. Schlitzer*) received **more than 2 millions SDN data files in ODV format and metadata file containing CDI information**. Data files were aggregated into a single T&S Data Collection using SDN Importer of ODV 4.5.3. This has been done in 9 pieces of about 250,000 files each, which have been re-combined at the end. The aggregation of the original temperature and salinity variables into single T and S variables was done using the Aggregated Derived Variables function. Logs files were sent to the coordinator and data centres for fixing eventual anomalies. During this phase more than 14000 files were rejected because ODV was not standard or not SDN standard. The list of errors was sent to 33 data centres and most of the data have been corrected. Finally regional sub-sets of TS data from 1900 to 2012 were created and distributed to Regional Coordinators (RCs) for QC (Quality Control).

3. Quality Assessment Process

Quality assessment process of regional T&S Collections started at the end of January 2013 by RCs. WP leader defined guidelines for a first basic QC analysis in ODV and a common template for the QC reports. In February 2013 sub sets of data 1990-2012 have been extracted from each TS Collection and released to MyO In-situ TAC in order to collaborate on the QC process, as defined during the First SDN-MyO Joint Meeting (18th of September 2012).

The outcome of the preliminary QC analysis was presented first by the WP leader to the **4th Steering Committee**, held in Paris on March 27th 2013, in order to share the work done and receive feedback and advices on the developments. A first shared decision between WP10 leader and the Steering Committee was that RCs should not modify data or Quality Flags (QF) but define procedures and priority actions in order to report on the data quality to CDI partners. This process might be applied iteratively to facilitate the update process and to progressively improve the overall quality of the infrastructure, consistently with what stated in SDN DoW. A second shared decision between WP10 leader and the Steering Committee was that Christine Coatanoan (Ifremer) with the support of Michele Fichaut (Ifremer) and Sissy Iona (HCMR) would coordinate the communication and the information flow between RCs-MyO INSTAC- NODCs.

An online **RC WebEX Meeting** was held on the 10th of April 2013 with the aim to share the work done, comment the results and find common strategies to feedback both the NODCs and MyOcean on the validation of the aggregated datasets. The WP leader presented also the Steering Committee outcome. The discussed topics were:

- the basic QC results on the raw aggregated data sets released to MyOcean INSTAC
- the coherency, coverage and quality of the data sets
- harmonization of the QC analyses conducted at regional level
- definition of the next data QC procedures
- the upcoming SDN-MyO Joint Meeting
- a possible update of the aggregated data sets

(see <http://www.seadatanet.org/Events/Products-meetings/WebEx-meeting> for more details)

The 15th of April RCs participated to the 2nd **SDN-MyO Joint Meeting** on common data quality assessment procedures and results. WP10 leader and MyOcean INSTAC leader presented a first feedback on the quality of the TS aggregated subsets and there was a subsequent discussion on the possible solutions and future steps. INSTAC partners had to generate feedback files according to the defined format by mid May with the target to provide the input to *Christine Coatanoan* and to allow a first feedback to NODCs (see for more details <http://www.seadatanet.org/Events/Joint-meetings/Second-SDN-MyO-meeting>).

The 27th of June another **WebEX Meeting** was organized by WP10 leader with RCs and the responsible persons of the WPs involved on the aggregation procedure to discuss about MyOcean feedback and the subsequent SDN work plan. *C. Coatanoan* presented a review analysis of all INSTAC anomalies and detailed how she organized the INSTAC anomalies per CDI partner. Figure 5 summarizes how it was organized by WP10 the feedback between MyOcean INSTAC and the CDI partners. It was decided to send in July to each CDI partner the corresponding list of anomalies asking to take actions in order to correct possible errors until mid September. Guidelines were also provided to explain the information in the anomalies files, how to make corrections on the data and how to send back resulting information in a detailed report (Figure 8) containing the updated anomalies (Figure 6) and CDI lists (Figure 7).

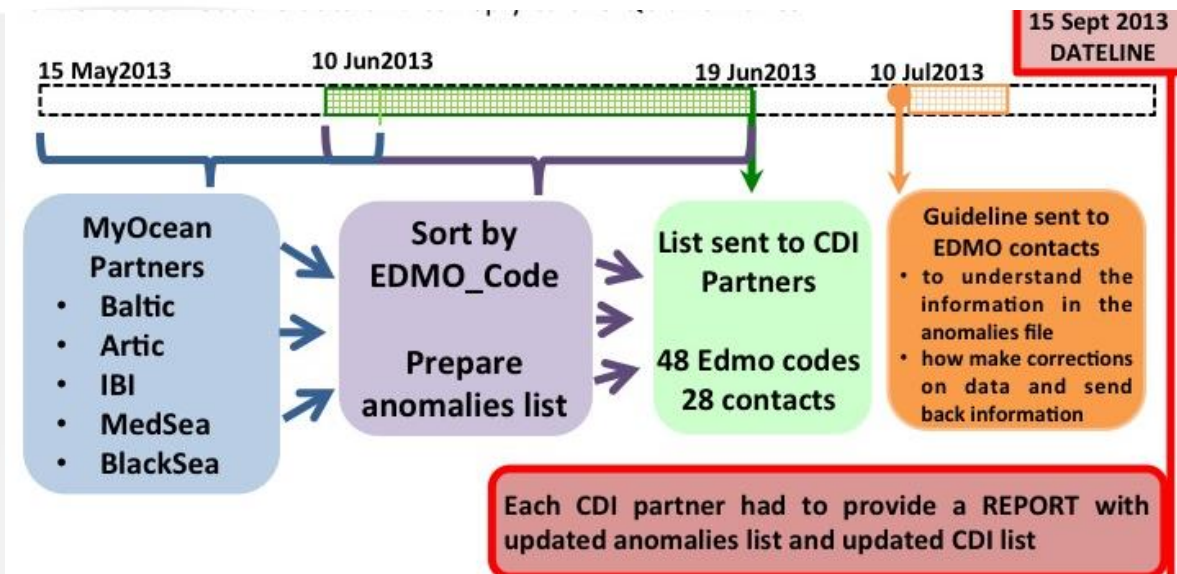


Figure 5 Timeline of the feedback procedure between MyOcean INSTAC quality assessment results on TS data collections and SDN CDI partners between June and September 2013.

Column	Description	Comment
LOCAL_CDI_ID	<i>cdi_identifier</i>	Partner local CDI identifier, Information from CDI
EDMO_CODE	EDMO_CODE of the organization distributing the data	Information from CDI
PLATFORM_CODE=CRUISE	CDI cruise_name	
STATION_DATE_START	Date at which the station starts	
STATION_DATE_STOP	Date at which the station ends	
UPDATE_DATE	Date of the control done by MyOcean partners	
PARAMETER	PARAMETER exported from ODV (TEMP, PSAL, DEPH or DEPTH [sometimes PRES when MyOcean partners have changed name])	
QC_ACTION	As described in Introduction, to define the type of anomalies (spike, gradient, missing value, etc)	
OLD_QC	QC from original dataset	
NEW_QC	QC suggested by MyOcean (see Annex I)	
VERTICAL_REFERENCE_START	Level at which starts the anomaly in the profile	
VERTICAL_REFERENCE_STOP	Level at which stops the anomaly in the profile	
AGREE WITH THE SUGGESTED CORRECTION (YES/NO)	Fill with Yes/No if you agree/disagree with the corrections suggested by MyOcean	
NODC COMMENT	Column to be added to your file in order to put some information about your our opinion about the suggested correction (agreement, disagreement, explanation if necessary)	
DETAILS	Column to be added to your file in order to put more information about suggested correction	

Figure 6 Table of the list of anomalies where CDI partners had to insert their comments.

LOCAL_CDI_ID	EDMO_CODE	PLATFORM_CODE=CRUISE
FI35199101301_00050_H10	486	PRIMO-0 21/03
FI35199443005_25900_H10	486	MBP-FRONT 1994
FI35199502002_00870_H10	486	EUROMARGE
FI35199706005_0K010_H10	486	PELMED 97
FI35199845001_00260_H10	486	BIODYPAR 1

Figure 7 Example of table of the list of updated CDI asked from CDI partners after data anomalies check procedure.

QC_Action	Number of anomalies detected by MyOcean	Number of true anomalies	%
Climatology	20	0	0.0
Gradient	328	1	0.3
IncreasingPressure	544	25	4.6
RegionalRange	42	0	0.0
Spike	148	42	18.3
StuckValue	28	0	0.0
VisualInspection	1	1	100.0
Total	1111	69	6.2

Figure 8 Template where CDI partners have to summarize the QC action, the anomalies detected by MyOcean INSTAC and the number of resulting true anomalies.

4. WP10 Quality Control

Six TS data collections, one per each European marginal sea (Arctic Sea, Baltic Sea, North Sea, North Atlantic Ocean, Mediterranean Sea, Black Sea), were analysed at regional level to assess and report on the quality of these products. The objectives were:

- to report to MyOcean about the quality of the regional collection sub-sets for the time period 1990-2012;
- to report to the NODCs about further and necessary improvements for the official data collection release (due for September 2013).

RC after a preliminary and basic QC through ODV software released the aggregated TS collections covering the time period 1990-2012, to the regional responsible of the INSTAC. The WP leader provided some guidelines and a report template in order to harmonize the work in progress.

The defined basic QC analysis steps were:

1. Station Selection Criteria: 1/Jan/1990-31/Dec/2012;

2. Polygon Selection in some regional basin (Mediterranean Sea) to discard data from some areas outside the analysis domain (Biscay Bay);
3. Data distribution and data density;
4. Histograms with annual and seasonal data distribution;
5. TS scatter plots of the entire dataset;
6. TS Scatter plots after the range check;
7. Scatter plot of observations with QF=1 (good), QF=2 (probably good);
8. Scatter plot observations with QF=0 (no quality check);
9. Statistics about QF
10. Visual control of scatter-plots to identify wrong profiles (outliers)
11. Visual check of spikes
12. Identification of stations falling on land
13. Identification of wrong or missing data

The scatter plots of the regional data sets highlighted the necessity of applying a gross range check, since there were observations with temperature and salinity out of reasonable values. Visual check of the scatter plots highlighted that some good data (QF=1,2) presented values out of ranges. Moreover there were a lot of data that did not pass any QC that looked reasonable. We supposed that some data centre probably inverted QF 1 and 0, thus we asked for checking. Outliers with respect to the defined ranges have been saved in text files in order to report to both MyOcean INSTAC and the NODCs.

RCs decided to improve the strategy for future QC analysis introducing specific sub-regional checks, per areas and per depth, and stability check on density.

5. MyOcean INSTAC Quality Control and feedback

The objective of MyOcean INSTAC of SDN and MyOcean collaboration is the integration of historical data from EuroGOOS ROOS partners and from SeaDataNet to serve operational oceanography users for multi-years ocean assessment (reanalysis). Due to the complex aggregation procedure, the SDN data delivery to MyOcean was a bit late with respect to the original time schedule. INSTAC focus was on aggregating the data from differences sources, ingesting the data from SDN rapidly and getting rid of the duplicates between the different data streams.

MyOcean INSTAC regional products will be built in two steps, as schematized in Figure 9:

1. MyOcean V3 product: aggregation of the data from ROOS providers and SeaDataNet National Data Centres, removing duplicates and converting all data in the same format with the same QC flags.
2. MyOcean V4 product: it will include SDN data after SDN scientific validation.

Figure 10 shows a schematic of INSTAC ftp branch dedicated to MyOcean V3 product of historical data, which nowadays contains SDN data. Figure 11 instead is a screenshot of MyOcean Interactive Catalogue describing the Mediterranean product and its data access (http://www.myocean.eu/web/69-myocean-interactive-catalogue.php/?option=com_csw&view=details&product_id=INSITU_MED_TS_REP_OBSERVATIONS_013_041). The other regional data collections can be accessed from MyOcean online catalogue.

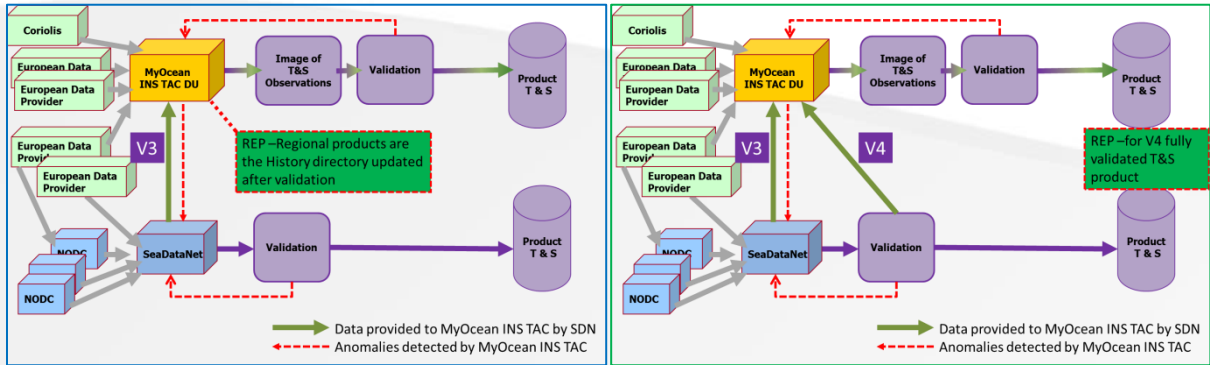


Figure 9 Building V3 and V4 MyOcean reprocessed T&S in situ products.

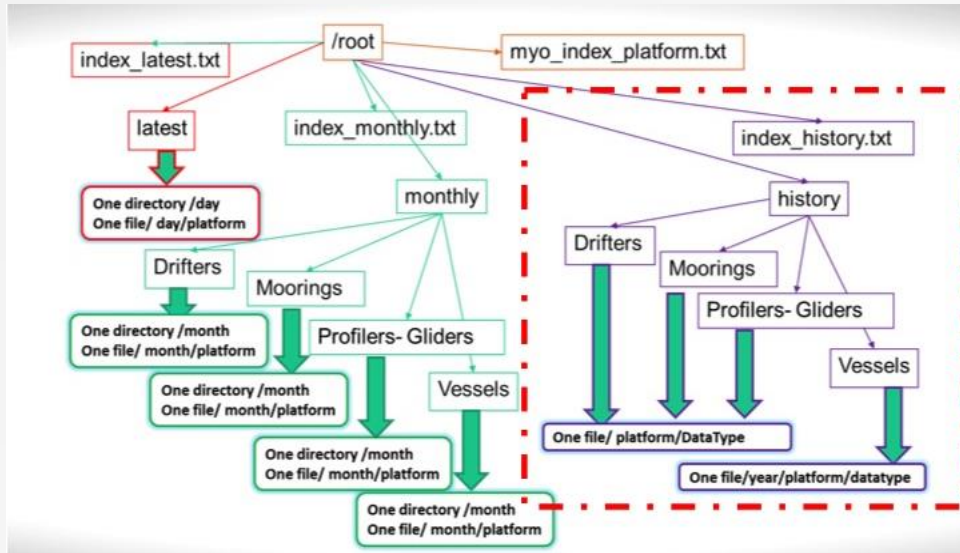
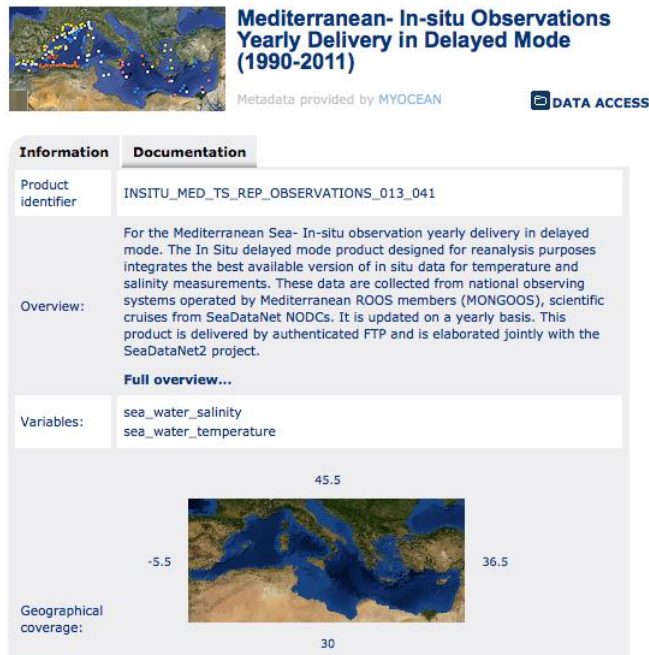



Figure 10 Schematic of the new branch history on MyOcean INSTAC ftp, which contains SDN data.

MYOCEAN INTERACTIVE CATALOGUE



Mediterranean- In-situ Observations Yearly Delivery in Delayed Mode (1990-2011)

Metadata provided by MYOCEAN  DATA ACCESS


Information	Documentation
Product identifier	INSITU_MED_TS_REP_OBSERVATIONS_013_041
Overview:	For the Mediterranean Sea- In-situ observation yearly delivery in delayed mode. The In Situ delayed mode product designed for reanalysis purposes integrates the best available version of in situ data for temperature and salinity measurements. These data are collected from national observing systems operated by Mediterranean ROOS members (MONGOOS), scientific cruises from SeaDataNet NODCs. It is updated on a yearly basis. This product is delivered by authenticated FTP and is elaborated jointly with the SeaDataNet2 project.
Variables:	<p>Full overview...</p> sea_water_salinity sea_water_temperature
Geographical coverage:	

Figure 11 Screenshot from MyOcean Interactive Catalogue describing the Mediterranean product and its data access.

MyOcean INSTAC regional groups applied their automatic quality check procedures (statistical test to check consistency over the area and climatological test to detect outliers) to SDN TS data collections and reported the issues causing problems for an easy aggregation with the rest of the data managed by the INSTAC. The main issue was that some important metadata have been lost in the aggregation process (i.e. Platform Code, Instrument Type for XBT). Therefore INSTAC partners decided to insert in their product only the data that were new and put aside the data that were detected as duplicates since it was difficult to know if these data were a better version of a data already managed by the INSTAC or another one. If INSTAC got a dataset directly from a ROOS partner, SDN version of it was discarded because it had less metadata.

The lack of Platform code is critical because all MyOcean data are organized by platform. The lack of WMO instrument type for XBT and the relative fall rate equation is an open question since this information was not recorded in the past.

INSTAC detected also some data with QF 0 (no control) mixed with QF 1 (good) either within a profile or in a series of profiles. Moreover they asked to know whether CTDs is calibrated or not since this could be a criterion to replace the same CDT received in Real Time (RT) with the one in Delay Mode (DM).

Globally the data were considered good but a visual QC was useful for checking suspicious data.

6. SDN analysis of INSTAC Anomalies

WP10 (C.Coatanoan) analysed the anomalies received from the regional groups of MyOcean INSTAC and summarized them in 8 categories:

1. **Climatology:** data falling out of the climatology envelope;
2. **Gradient:** gradient anomaly within the profile;
3. **Increasing Pressure:** pressure must increase monotonically along the profile;
4. **Missing Value:** missing value but QF not equal to 9;
5. **Regional Range:** value falling out regional range;
6. **Spike;**
7. **Stuck Value:** constant profile;
8. **Visual Inspection:** data anomaly detected visually.

Figure 12 presents a graph with the distribution of the different categories of anomalies from which appears maximum percentages of “Increasing Pressure” and “Missing Values” anomalies.

It was found that “Missing Value” 0 to 9 came from a bug in ODV during data export thus it could not be considered a true anomaly. The ODV bug has been fixed.

While “Increasing Pressure” anomalies could have several origins:

- In case of bottle platform it could be normal to have several samplings at the same depth. In this case there is no anomaly.
- It is a duplicate CDI problem when data providers forwarded multiple updates of the same CDI in a short time frame with overlaps. Also in this case it is not a true anomaly.
- It is a true error of non-increasing depth/pressure and then it should be flagged to 4.

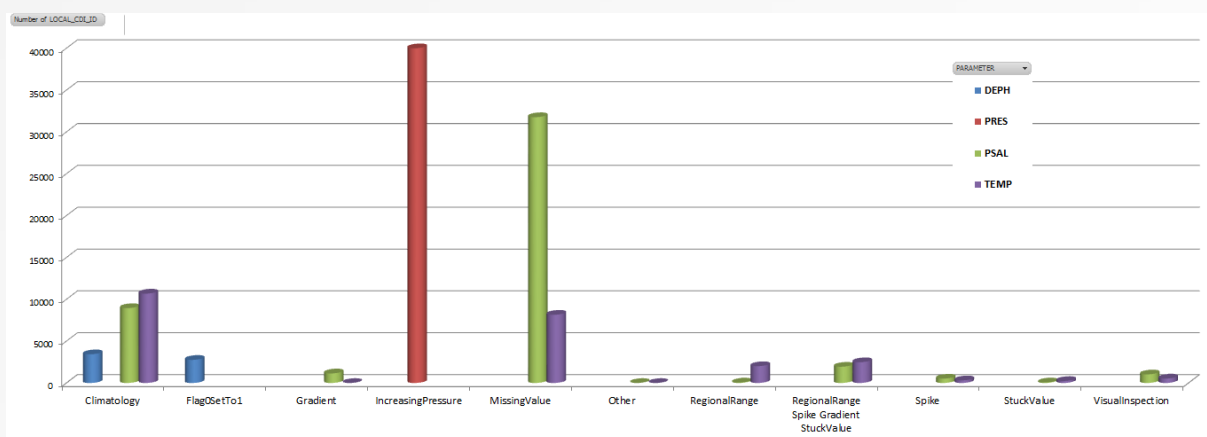


Figure 12 Outline of MyOcean type of anomalies as analyzed by WP10 (by C.Coatanoan).

7. Results of the first aggregation and quality assessment

Data Aggregation Procedure was an extensive and fruitful exercise to manage more than 1 Million data and involving many WPs, people and institutions. It was huge distributed effort involving 62 data centres and more than 300 data originators.

The **Quality Assessment Procedure** took place in coordination between SDN RCs, MyOcean INSTAC, SDN CDI partners and permitted to identify and correct lots of data. **Aggregation and Quality assessment** allowed to **ameliorate and refine each technical phase** but mainly **to highly improve the quality of SDN infrastructure content**. The collaboration between SDN and MyO was crucial during the data quality assessment and allowed **to identify and correct lots of data anomalies**. WP10 promoted collaborations and communication between partners, WPs (WebEX conf) and projects (2 Joint Meetings). All RCs were active and collaborative in QC assessment and reporting activities during the second year of SDN Project.

Since the quality of historical data collections would have been highly improved by an update before the official release, we decided during the WebEX Meeting on products (June 27th) to repeat the aggregation procedure and consequently the QC assessment (see time schedule in Figure 3). A new aggregation exercise was set up during the 5th Steering Committee in September 2013 in order to deliver the best aggregated data sets. A new exercise would enable also to retrieve **restricted data** to be used for the statistical products generation (climatologies, maps, profiles). In this way SDN products, based on updated data collections, will be higher quality.

8. Second data aggregation and quality assessment analysis

Between November 2013 and March 2014 a new harvesting procedure was implemented to retrieve the data within SDN data base and a new aggregation of data into ODV collection was performed, according with the implemented QC strategy described in Figure 1.

First version of the V1.1 aggregated dataset has been provided to WP10 by AWI at the end of January 2014. Some ODV bugs have been detected (mix of several profiles in a same profile, gap in metadata rows) working on the new data sets and have been fixed by a new version of ODV. A new version of aggregated dataset was delivered too in March 2014. Another ODV improvement is also the new parameter Pressure in addition to Depth to identify levels.

It follows the list of ODV bug fixing activities, done in collaboration between some WP10 RCs (Ifremer, SMHI) and AWI, that contributed to better release of ODV version:

- 24 Jan: AWI provided the first version of aggregated data set (ODV 4.5.7);
- Ifremer split into regional data sets;
- 29 Jan: new release ODV 4.6.0;
- 4 Feb: Feedback on ODV 4.6.0 about problems with log file (no way to record changes on QC) -> Ifremer used ODV 4.5.7 to edit log;
- 6 Feb: Feedback about ODV bug (Temp and TEMP were not unique labels) that was fixed with ODV 4.6.0. However the second dataset was created with ODV 4.5.7;
- 13 Feb: new version ODV 4.6.1 to test (Export stations history to get changes on QC);
- 18 Feb: official version ODV 4.6.1;
- 20 Feb: SMHI found a problem of ODV 4.6.1 with Metadata;
- 25 Feb: problem with salinity (3 data files being merged into 1 profile);

- 27 Feb: gaps in temperature profile levels;
- Problems between versions due to creation in ODVCF5 or ODVCF6 format collections (ODVCF5 = no history feature, all edit logs are written to the collection log file);
- 3 Mar: AWI proposed a new aggregated dataset and a new release of ODV 4.6.2;
- 10 Mar: a new aggregated dataset was released to WP10. Collections are ODVCF5 and can be opened by ODV 4.5.7 as well as ODV 4.6.0 or later;
- 17 Mar: new regional data collections and restricted data sets provided to WP10.

9. Regional Aggregated Data Sets

It follows the description of the regional data collections.

9.1. Mediterranean Sea

9.1.1. General Characteristics of the Mediterranean Sea historical data set

The historical data collection of the Mediterranean Sea contains Temperature and Salinity observations between -9 and 37 degrees of longitude, thus including an Atlantic box and the Marmara Sea that will be treated separately. The spatial distribution and the data density of Temperature and Salinity observations from the entire data collection are shown in Figure 13 a and b. The spatial distribution of data (Figure 13a) presents a good data coverage in the Western Mediterranean basin and the Atlantic box, while in the Eastern Mediterranean many areas are characterized by few and sparse data, like the coastal areas Tunisia, Libya, Croatia and Turkey. Data density map (Figure 13b) highlights that observations are more concentrated along the coastal areas of Spain, France and Italy (Ligurian Sea and Northern Adriatic Sea). In the eastern part of the Basin, maximum data concentration is along the Israeli and the Greek coasts.

Temporal distribution of data is in Figure 14. Annual distribution (a) proves that data are very sparse before 1945 and they start to increase systematically from the sixties and concentrate mostly between the end of the nineties and beginning of noughties. The highest peak corresponds to year 2009, when a lot of sea gliders observations have been sampled along the coasts of France and Spain. This must be taken into consideration during climatological data analysis, like climatology computation. Seasonal distribution of data present a good coverage all year long but a peak is present at the end of summer beginning of autumn (September, October) and this might be due to surveys dedicated to monitor particular events. Further investigations are needed for climatological analysis or other applications.

Figure 15 show Temperature (a) and Salinity (b) data distribution. Salinity observations are less and sparser than temperature ones. Both maps show the presence of data along ship tracks, along coastal transects, and regular monitoring arrays. Tab. 1 condenses the number of observed stations and its repartition in Temperature stations and Salinity stations and stations that sampled both T and S.

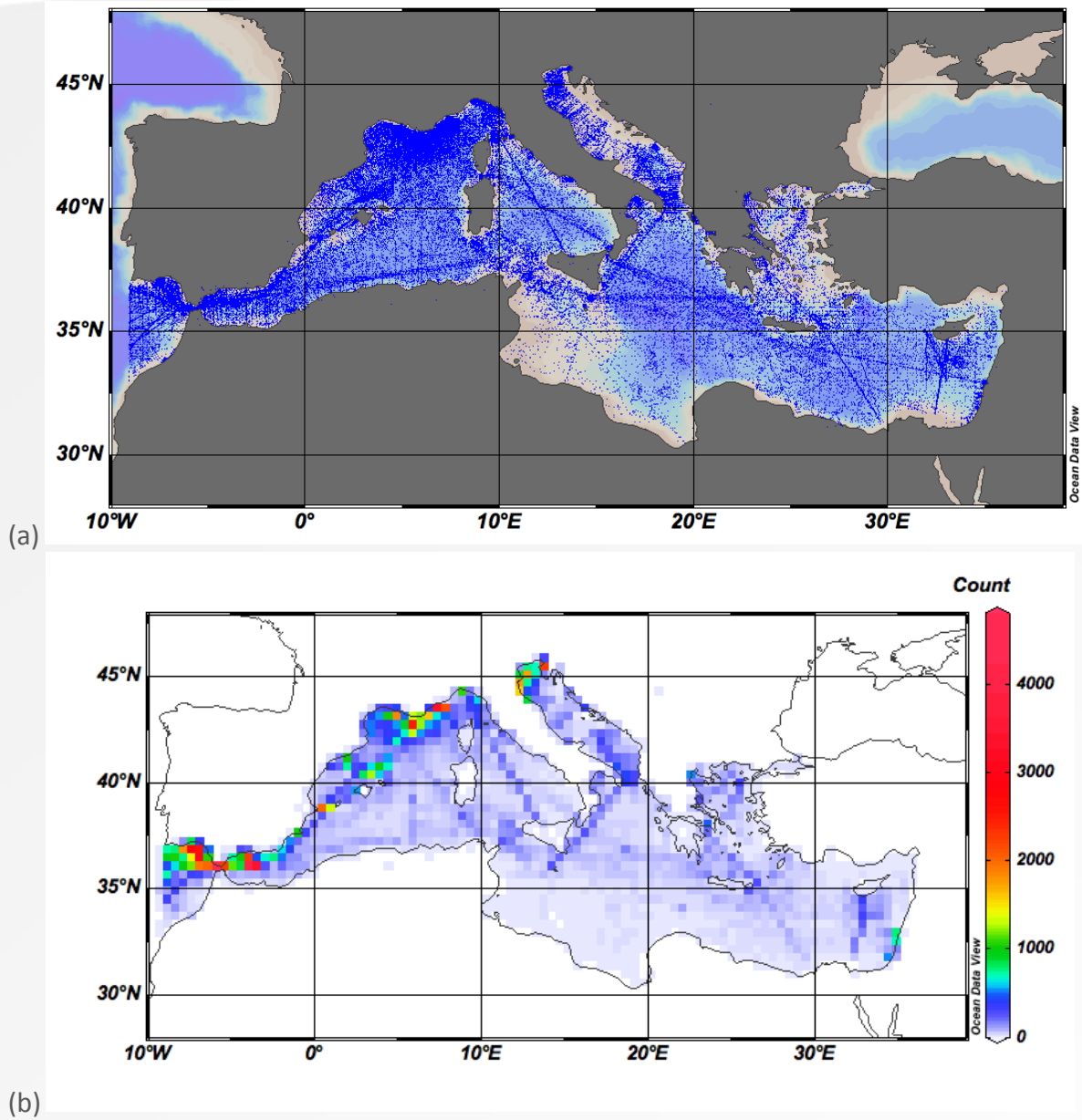


Figure 13 Temperature and Salinity data collection for the Mediterranean Sea in the time period 1900-2012: (a) Data distribution map; (b) Data density map.

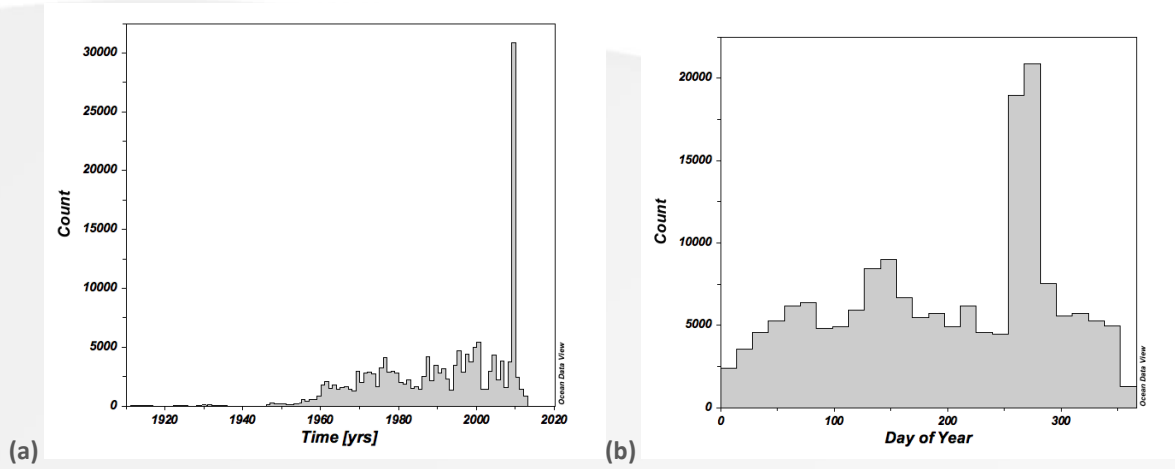


Figure 14 (a) Annual data distribution and (b) seasonal data distribution for the time period 1900-2012 in the Mediterranean Sea.

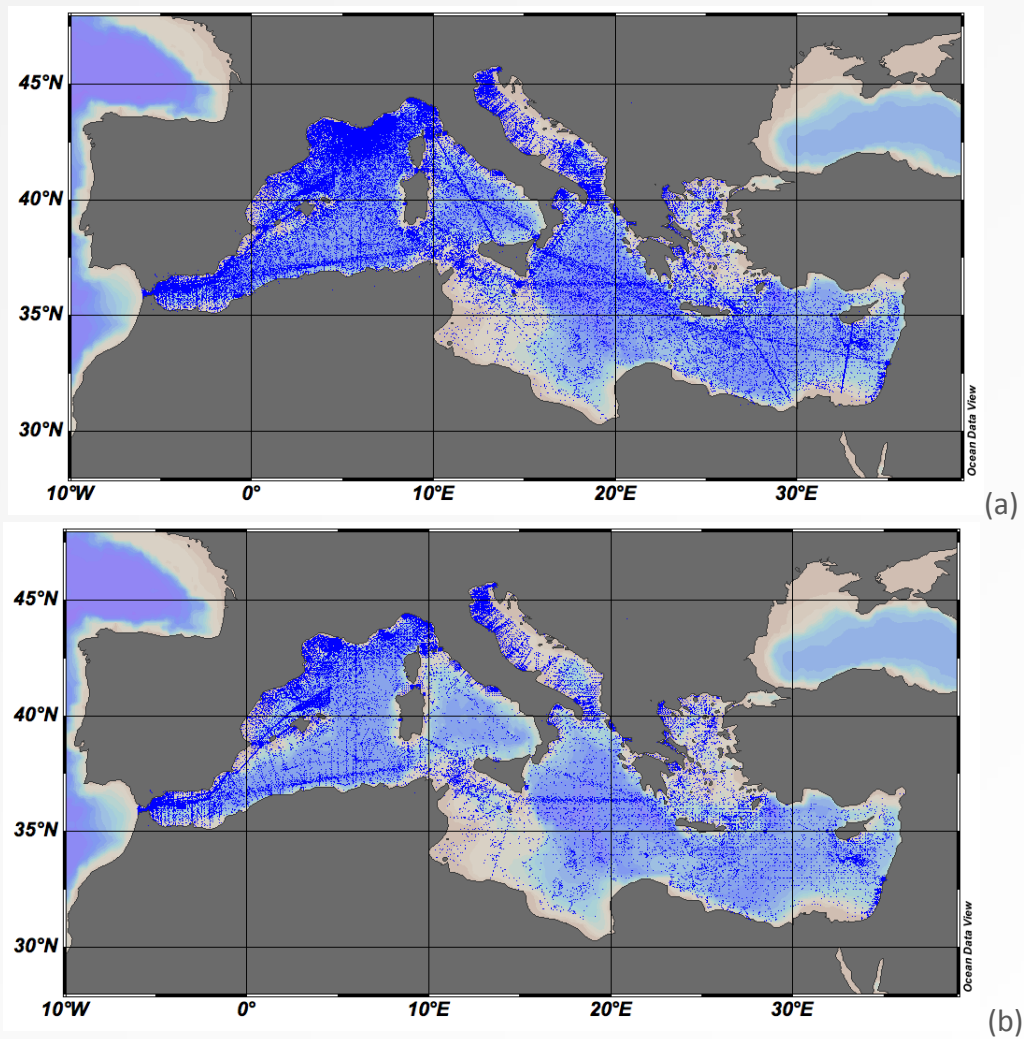


Figure 15 Data distribution map for the Mediterranean Sea: (a) Temperature, (b) Salinity.

PARAMETER	STATIONS
total	169438
T	165243
S	110670
TS	109249

Tab. 1 Number of data points for Temperature, Salinity and TS couples For the Mediterranean Sea only.

After a general description of the historical data set a visual control of all observations allowed to assess their quality and to identify the principal criticalities for possible future applications and users.

Temperature scatter plots are presented in Figure 16. In particular panel (a) represents all T observations with QF equal to 1, from which some bad observations could be easily identified. Panel (b) represent all temperature observations with QF equal to 0 that did not pass through a QC analysis from data providers. These data seem good data and could be analysed in order to use them in the future.

Tab. 2 includes the repartition of observations according to their QF and from which it comes out that good temperature data correspond to the 96.8% of the total, while data still not checked are 2.9%.

Salinity scatter plot of good data (Figure 17a) confirms that there are some bad ($S > 60$ psu) data flagged as good. There are also a lot of coastal data with salinities close to zero. The user might keep this in mind for applications like climatological map computation that could present strong or unrealistic gradients along the coastline. The scatter plot of salinity data not quality checked (Figure 17b) reveal the presence of good data that could be recovered and analysed. Good data are a 95.3% while data with QC=0 represent a 4.5% (see Tab. 2).

	TOT	QF0	QF1	QF2	QF>3
T	24889496	718157	24086659	49610	35070
		2.9%	96.8%	0.2%	0.1%
S	16121430	718517	15360727	0	42186
		4.5%	95.3%		0.3%

Tab. 2 Number of data points for Temperature and Salinity and their subdivision according to Quality Flags (QF) 0, 1, 2 and from 3 to 9.

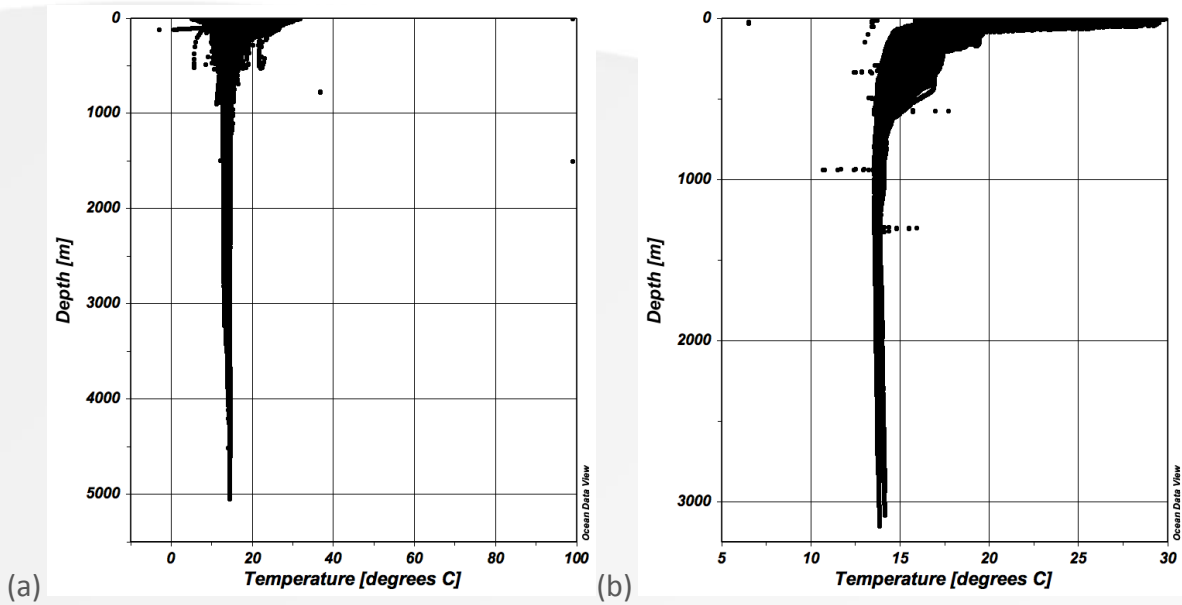


Figure 16 Temperature scatter plots: (a) Data with QF1; (b) data with QF0.

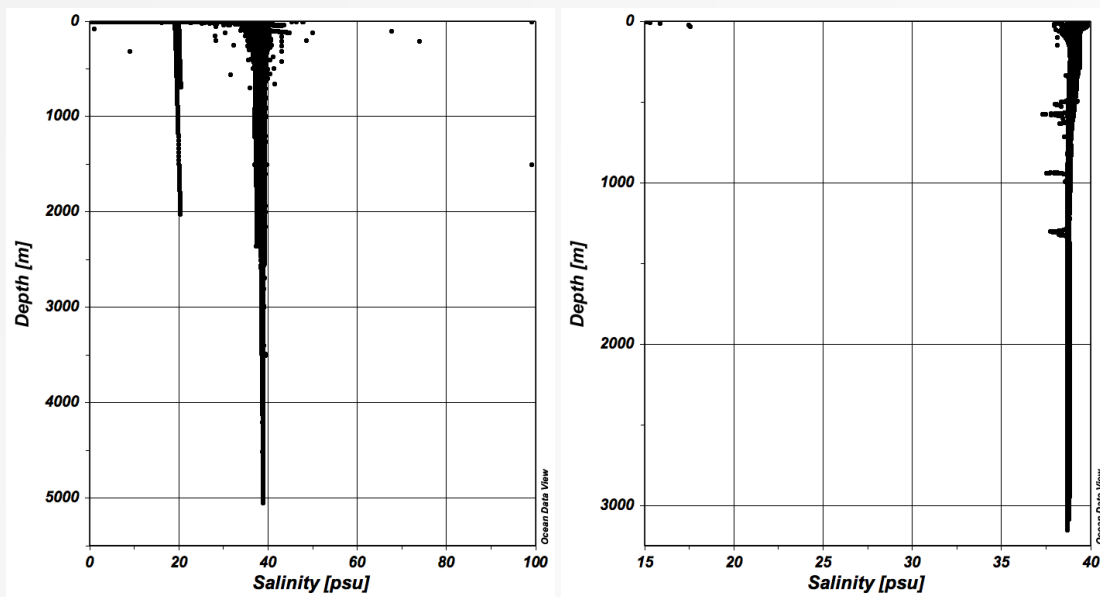


Figure 17 Salinity scatter plots: (a) data with QF1; (b) data with QF0.

Another way to visual check a data collection of temperature and salinity observations is to look at TS scatter plots. Figure 18 present the entire data collection in (a) while (b) includes only data with QC equal to 1 and (c) contains data not checked (QF=0). It is evident that some bad data passed the QC procedure and they have been erroneously flagged 1. Plot (c) confirms that most of the data not checked could be analysed and recovered for future usage.

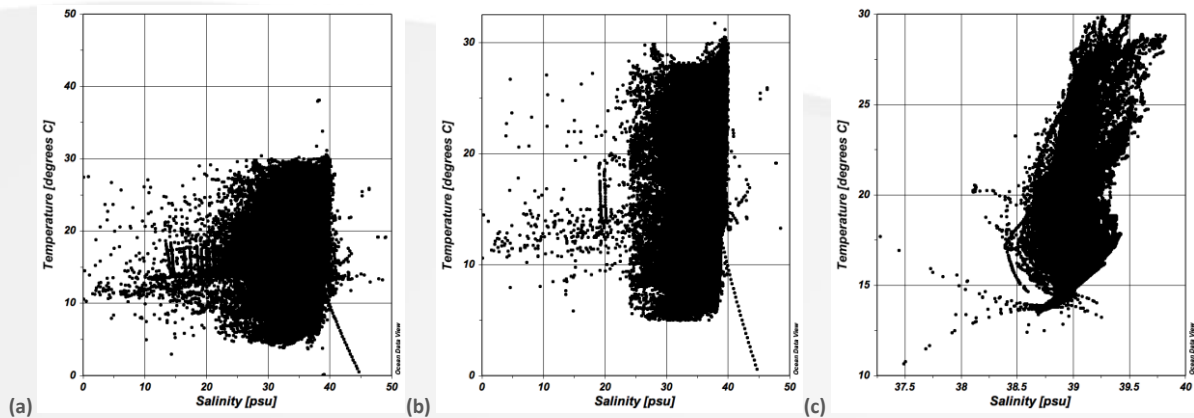


Figure 18 Mediterranean TS data collection for the time period 1900-2012: (a) TS diagram of full data collection; (b) TS diagram considering only data with QF1 (good) and QF2 (probably good); (c) TS diagram considering only data with QF0 (no quality control).

Another quality issue, concerning the Mediterranean Sea data collection, is the presence of observation with obvious wrong locations on land, especially in the southern part of Spain, Italy and the coast of Morocco. Part of observations on land might be coastal observations that appear on land due to the low resolution land mask used by ODV, thus a list of observations on land have been extracted and it will be sent to NODCs asking to check the right positioning.

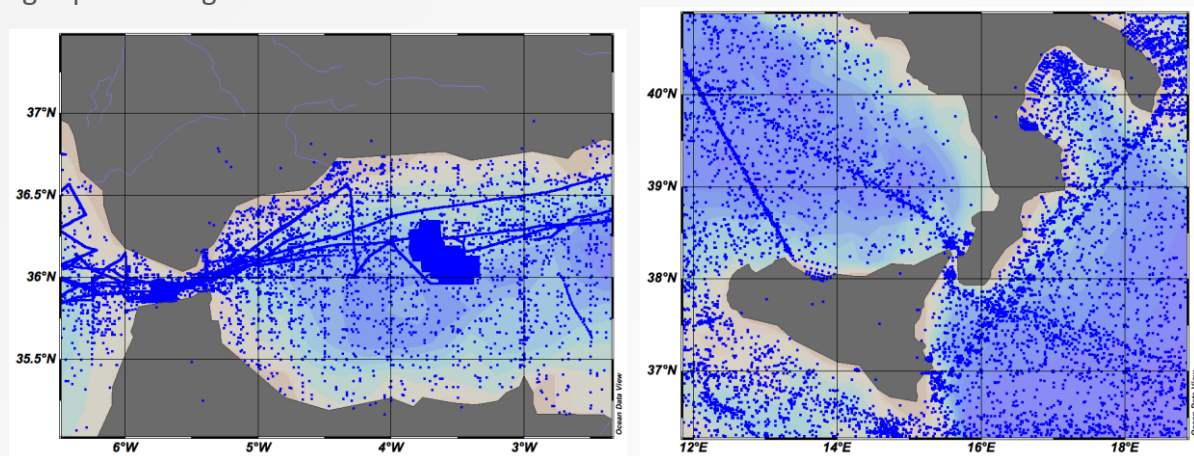


Figure 19 Zooms from the Mediterranean Sea data distribution map that unveil the presence of observations with erroneous locations on land.

9.1.2. Atlantic Box

The Mediterranean Sea data collection contains also an Atlantic box defined between -9 and -6 degrees of longitude. This area has been included in order to have an overlapping zone with the North Atlantic region. The data set consists of 31322 stations, mostly concentrated along the Spanish coastal areas (see Figure 20).

The temporal distribution of the Atlantic data in Figure 21 present the same peaks of Figure 14 for the Mediterranean data and this might be due to the presence of a lot of sea gliders observations sampled in 2009.

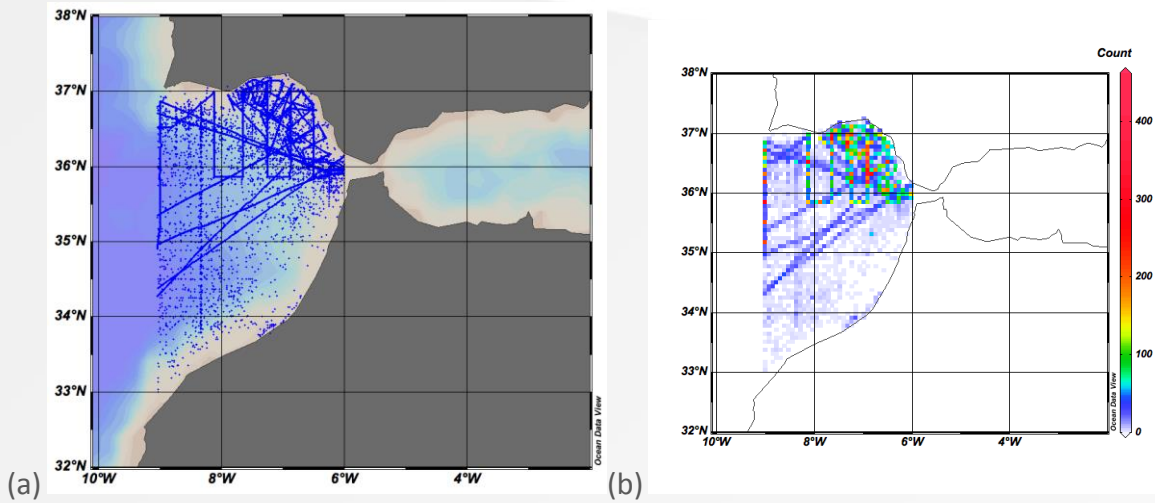


Figure 20 Data distribution map (a) and data density map for the Atlantic box included in the Mediterranean historical data collection. (31322 stations)

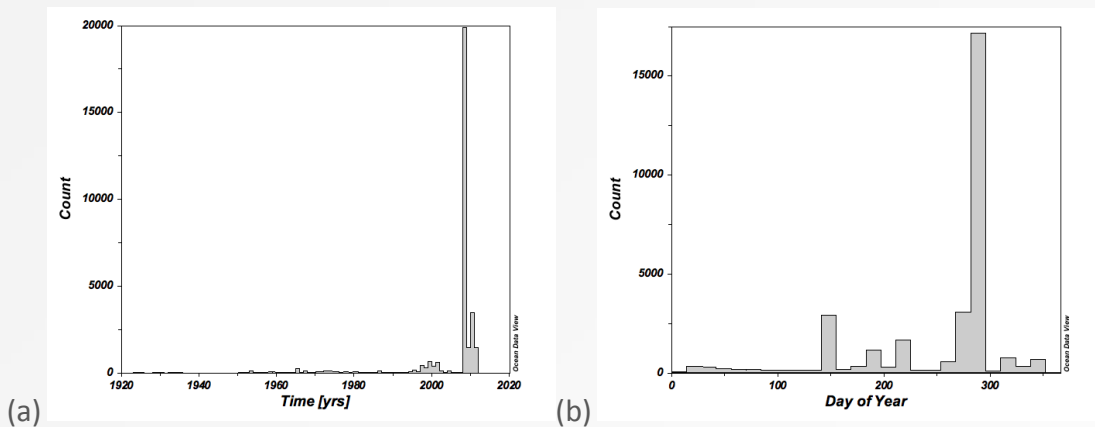


Figure 21 Annual (a) and seasonal (b) data distribution in the Atlantic box.

Figure 22 shows Temperature scatter plots of the entire data set (a) and of data flagged as good (b) and not checked data (c). Visual check did not reveal anomalies since some bad data were filtered out when QF=1 was selected. Data without QC (c) represent a single profile with 1017 observations along the water column.

Salinity scatter plots in Figure 23 display some anomalous observation (a) that in part are filtered out when QF=1 is selected (b) but some data with very low (<20psu) salinity should to be further analysed. Salinity observations that did not pass through a QC (c) belong to a single profile.

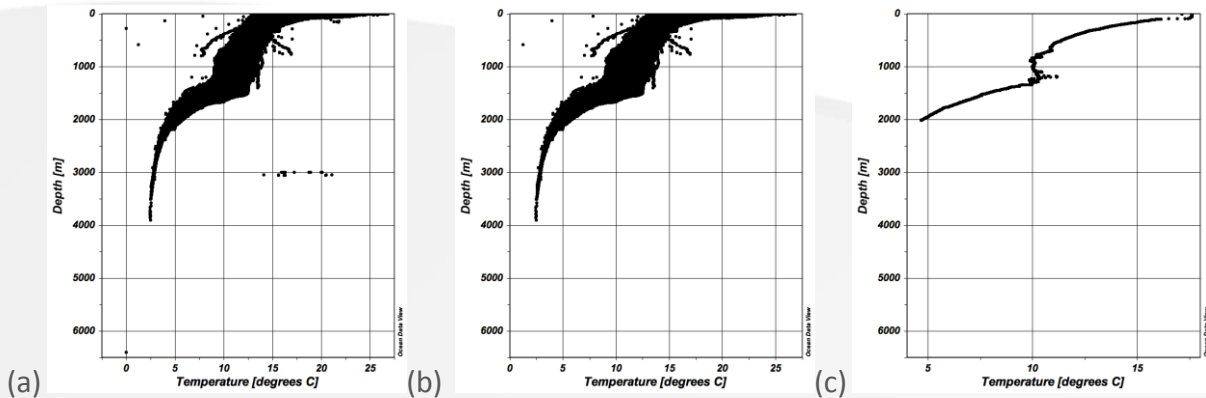


Figure 22 Temperature scatter plots for the Atlantic box: (a) entire data set; (b) data with QF=1; (c) data with QF=0.

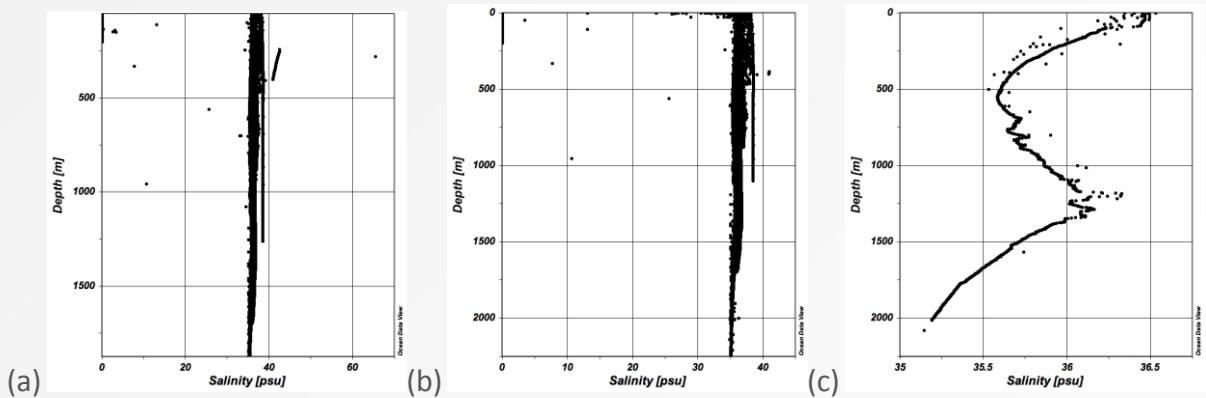


Figure 23 Salinity scatter plots for the Atlantic box: (a) entire data set; (b) data with QF=1; (c) data with QF=0.

	TOT	QF0	QF1	QF2	QF>3
T	2495982	1017	2494121	5	839
S	552260	1076	550291	0	893

Tab. 3 Number of data points for Temperature and Salinity data within the Atlantic box in Figure 20 and their subdivision according to Quality Flags (QF) 0, 1, 2 and from 3 to 9.

9.1.3. Marmara Sea

The Marmara Sea has been defined between 26 and 30 longitude degrees and 39.5 and 41.5 of latitude. The data set consists of 156 stations (Figure 24). Observations are few and very sparse both in space and time, as shown in Figure 25. This makes unfeasible the computation of reliable statistical products, thus we will encourage data providers to release more observations or we will look for additional data sets.

Figure 26 presents Temperature scatter plot for the entire data set (a) and for observations with QF=1. Very few points (19) were filtered out from QC as it could be checked in Tab. 4. Figure 27 shows the Salinity scatter plots for the entire data set (a) and for good observations. Few observations (55) were excluded from good data.

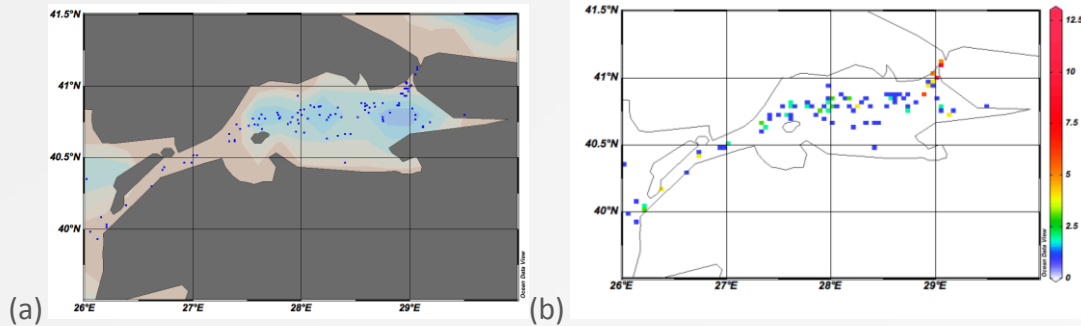


Figure 24 (a) Data distribution map and (b) data density map for the Marmara Sea.

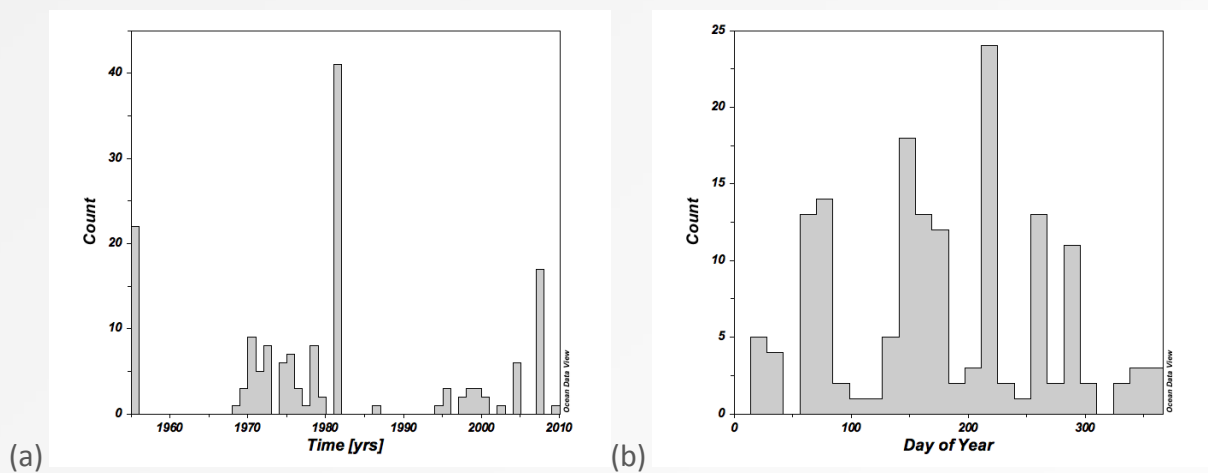


Figure 25 Annual (a) and seasonal (b) data distribution in the Marmara Sea.

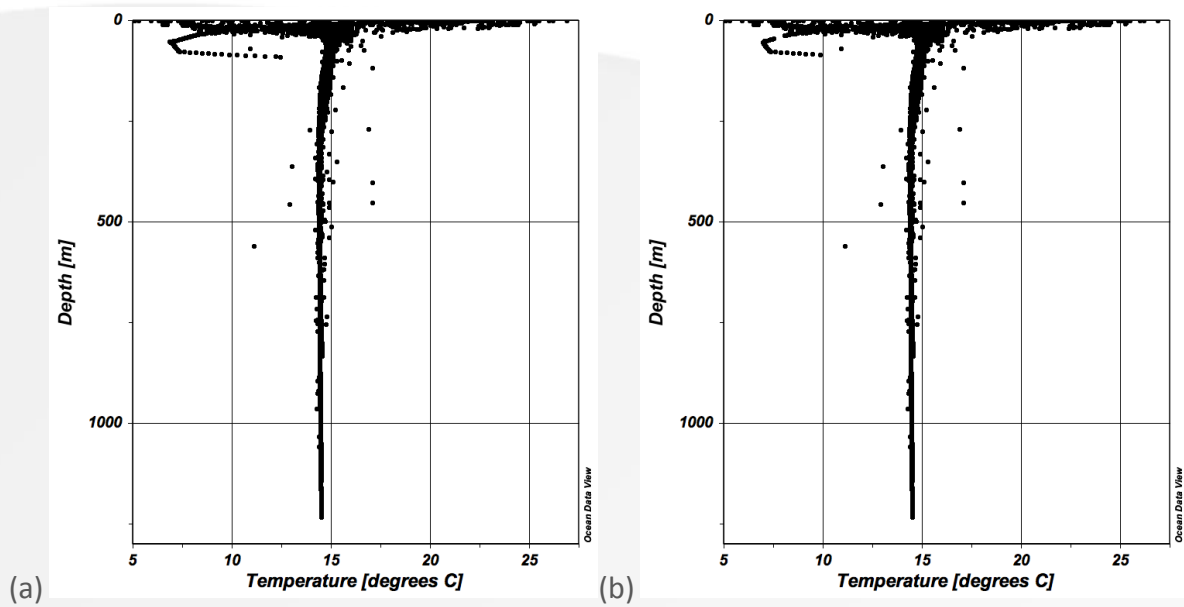


Figure 26 Temperature scatter plots for the Marmara Sea data set: (a) entire data set; (b) data with QF=1.

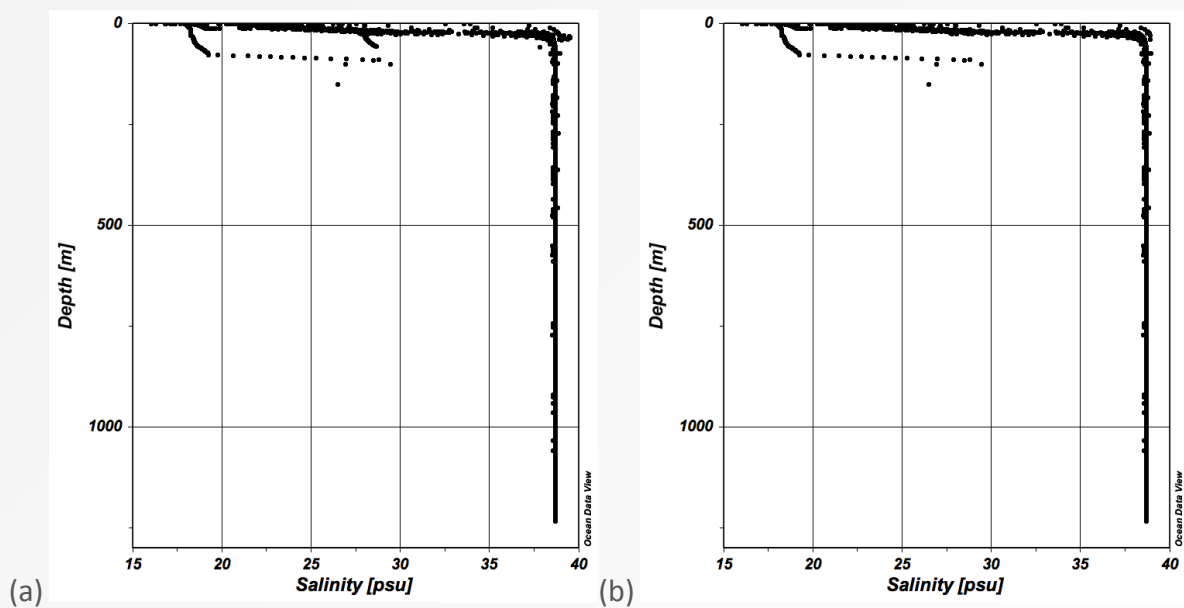


Figure 27 Salinity scatter plots for the Marmara Sea data set: (a) entire data set; (b) data with QF=1.

	TOT	QF0	QF1	QF2	QF>3
T	17979	0	17940	20	19
S	16223	0	16168	0	55

Tab. 4 Number of data points for Temperature and Salinity data within the Marmara Sea in Figure 24 and their subdivision according to Quality Flags (QF) 0, 1, 2 and from 3 to 9.

9.1.4. The Mediterranean Sea restricted data collection

Restricted data collection for the Mediterranean Sea, retrieved during last aggregation exercise, represent 17% of the total number of data (28690 stations) retrieved during the aggregation procedure. This is still a too high percentage of data and we estimate that another 20% of data were not retrieved from the infrastructure because CDI partners did not approve the request. We will encourage data providers to adopt a data policy of free and open access in provision of data to users, in line with international agreements (e.g. WMO, IOC, ICSU, GEO/GEOSS). Figure 28 show the data distribution (a) and density maps (b) from which we might notice a lot of data in the Adriatic Sea, the Sicily Channel and the Tunisian coast that are crucial for computing good quality statistical products. The temporal distribution Figure 28c indicates that restricted data concentrate mainly in the time period after 1995.

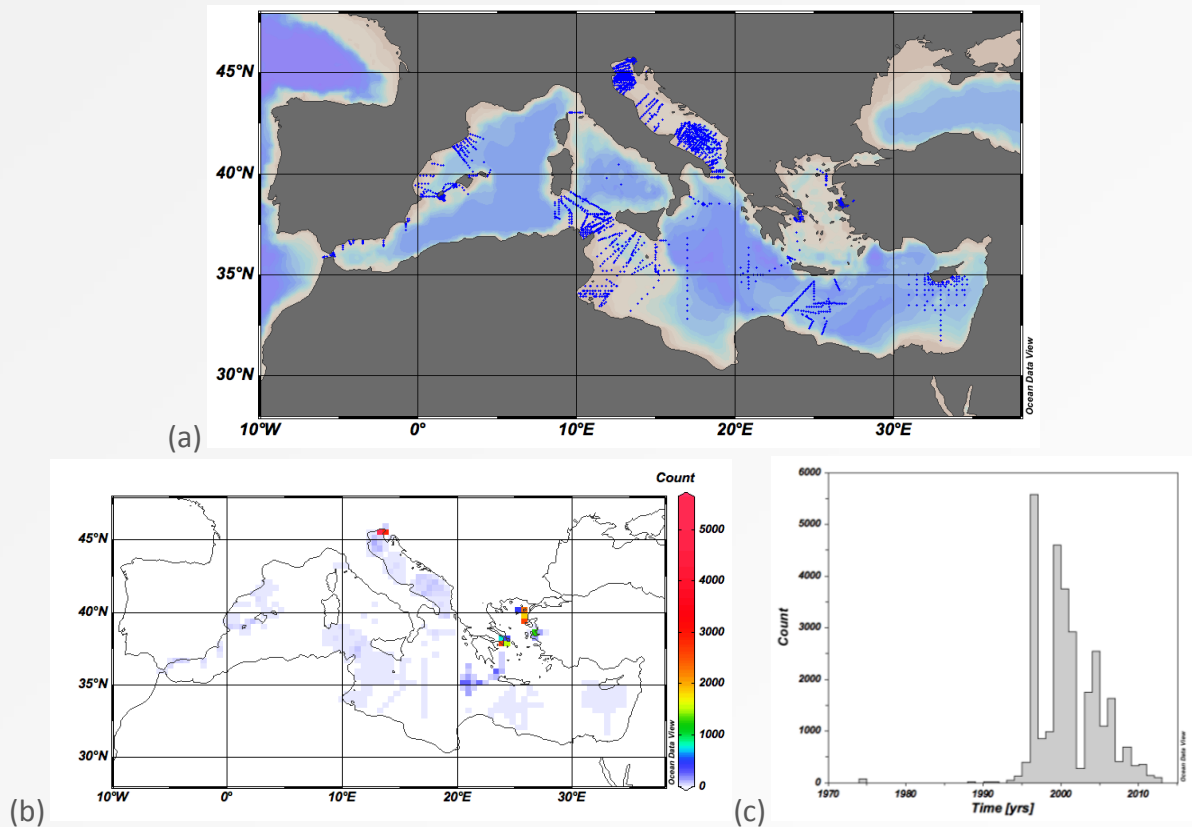


Figure 28 The Mediterranean Sea restricted data collection: (a) Data distribution map; (b) Data density map; (c) annual distribution map.

	TOT	QF0	QF1	QF2	QF>3
T	6638252	2094	6621344 (98.8%)	0	11414 (0.2%)
S	6622804	67791 (1%)	6542113 (98.8%)	0	12900 (0.2%)

Tab. 5 Number of Temperature and Salinity data points for the Mediterranean Sea restricted data collection and their subdivision according to Quality Flags (QF) 0, 1, 2 and from 3 to 9.

9.2. Black Sea

9.2.1. General Characteristics of the Black Sea historical data set

The Black Sea historical dataset includes data from the Black Sea and Sea of Azov for period 1868 – 2013. The initial number of stations is 125565. However the initial dataset contained significant number of duplicates (22500) and empty stations, which were excluded in process of quality control. The remaining number of stations is 101604. From those 2379 stations appear to be on land as per ETOPO1_2 min topography. The statistics below are obtained for the Black Sea Dataset cleaned from duplicates, empty stations and stations on land.

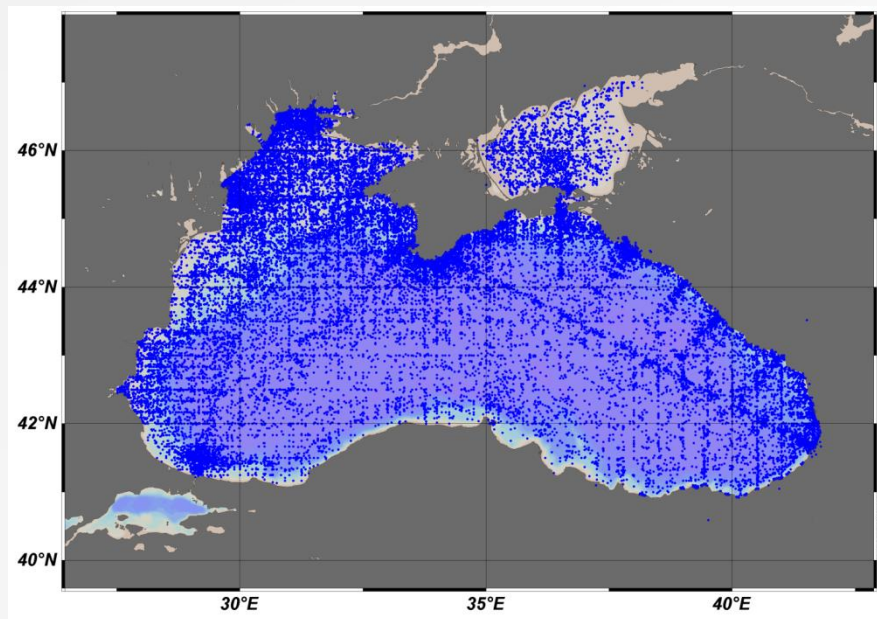


Figure 29 Data distribution map: Temperature at the surface (98870 stations)

Black Sea public data	2013	2014	Increase %
All station	118488	125565	6.0
Without duplicates and empty stations	99530	101604	2.1

Tab. 6 Number of stations in the public Black sea data set.

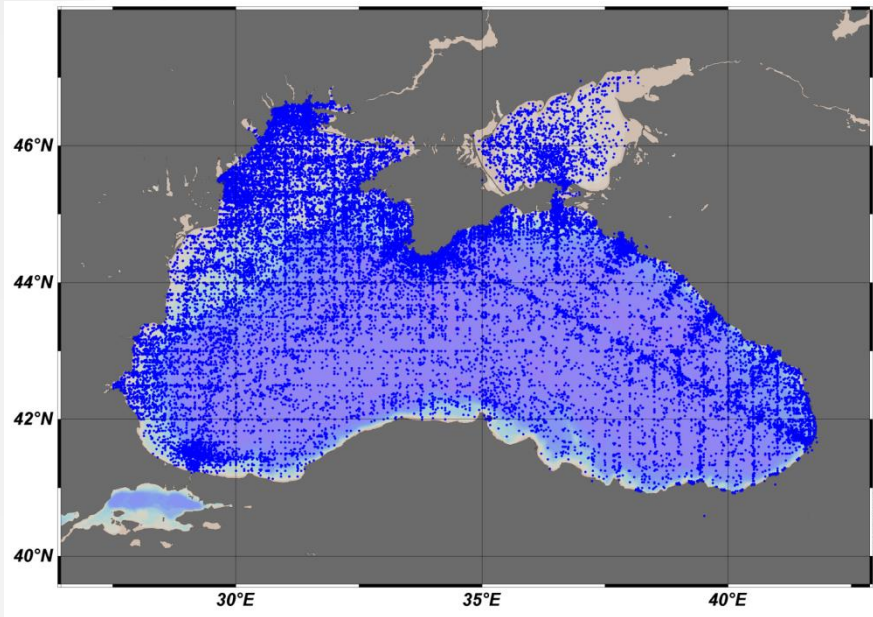


Figure 30 Data distribution map: Salinity at the surface (93688 stations)

The data distribution maps (Figure 29 for Temperature and Figure 30 for Salinity) allows identifying areas of intensive observations such as NW of Black Sea, South of Crimea, near Bosphorus area, areas along standard transects etc., however the interior of the sea and its southern part are rather poorly covered with observations. There is no significant difference between Temperature and Salinity distribution maps.

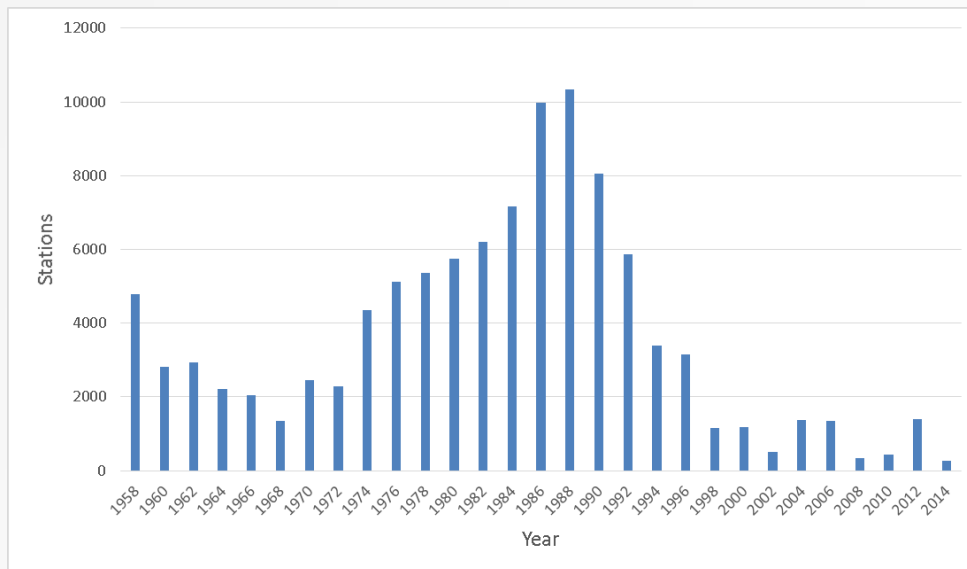


Figure 31 Annual distribution of stations with Temperature observations (left column represents all stations before 1958)

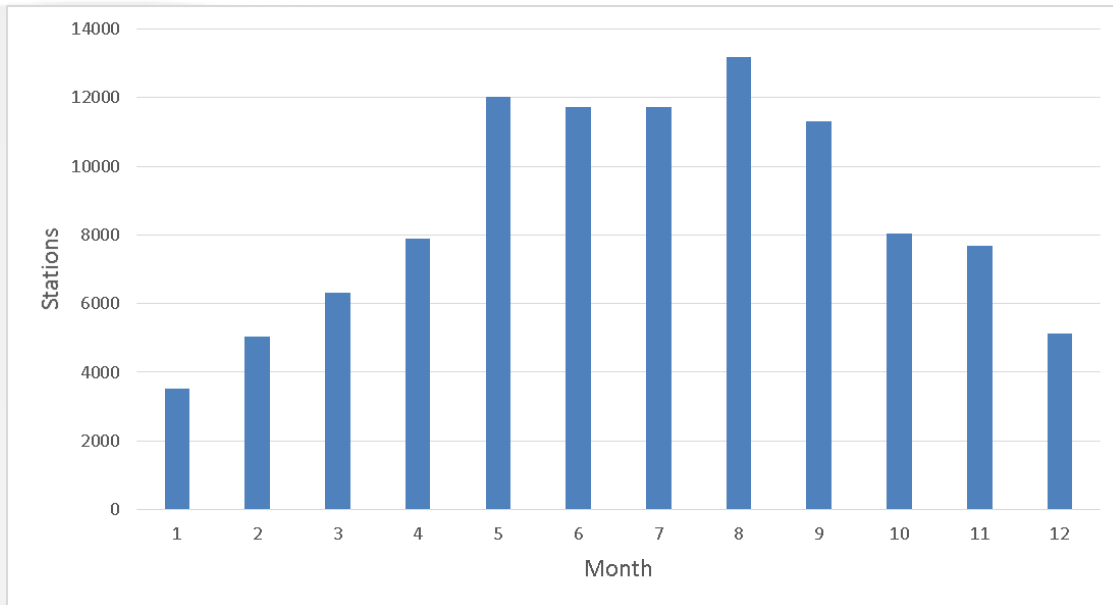


Figure 32 Monthly distribution of stations with Temperature observations

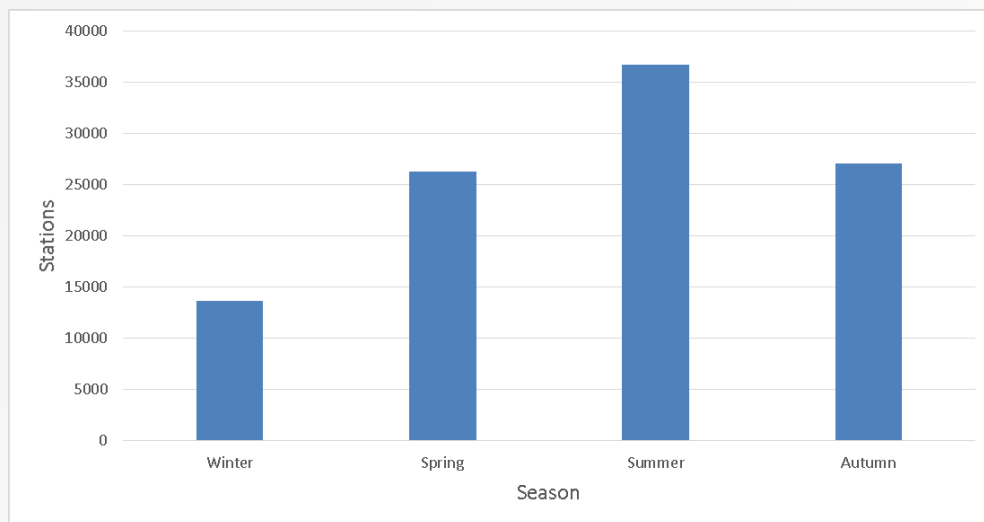


Figure 33 Seasonal distribution of stations with Temperature observations

The annual (Figure 31), monthly (Figure 32) and seasonal (Figure 33) distributions of Salinity observations are similar to those of Temperature. The most intensive oceanographic observations were performed in Black Sea in period 1970 – 1995. Most of observations were performed during summer months while in winter period intensity of observations is almost 3 times less than in summer.

The availability of data drastically decreases with depth as shown in Figure 34.

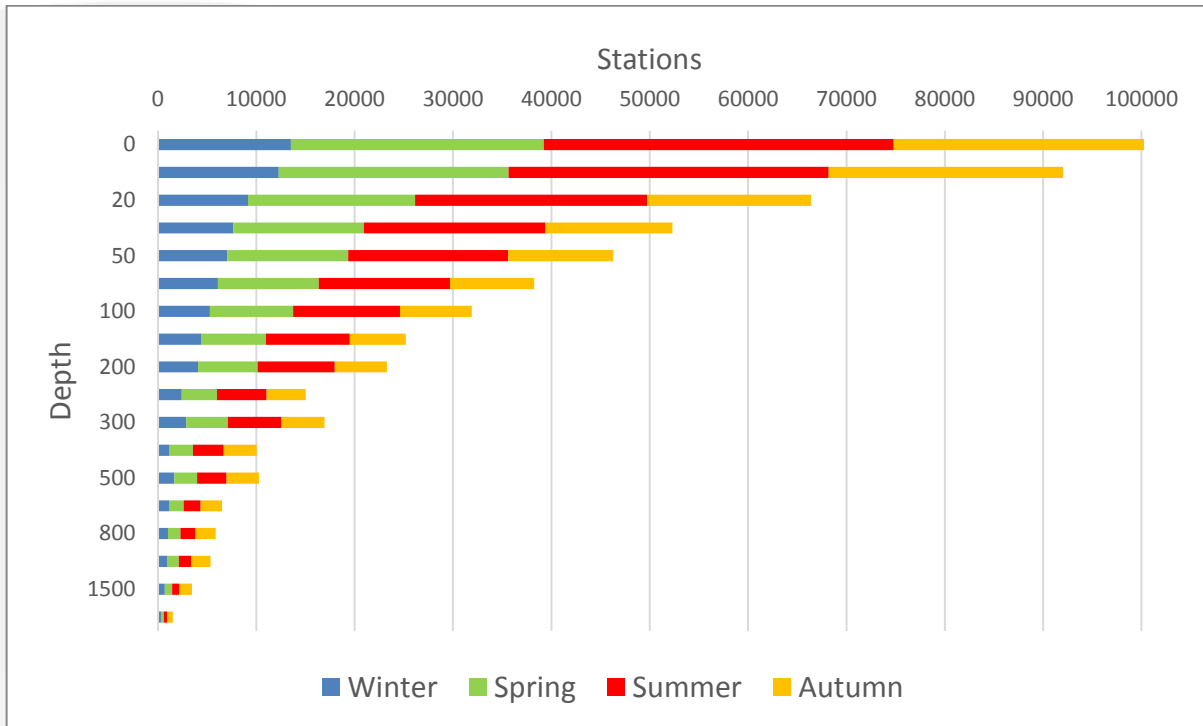


Figure 34 Vertical distribution of Temperature data along the IODE standard levels

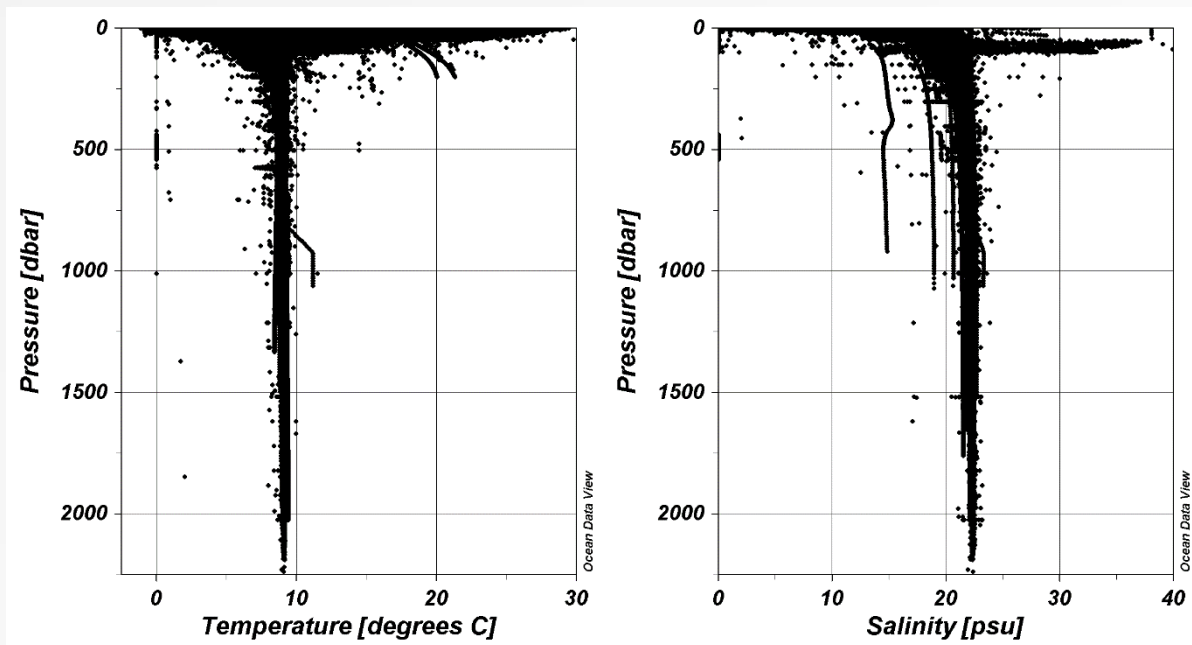


Figure 35 Temperature and Salinity scatter plots for entire dataset before QC

Temperature and Salinity (Figure 35) scatter plots before QC demonstrate presence of wrong profiles and outliers, which in many cases however are flagged with QC=1 (good).

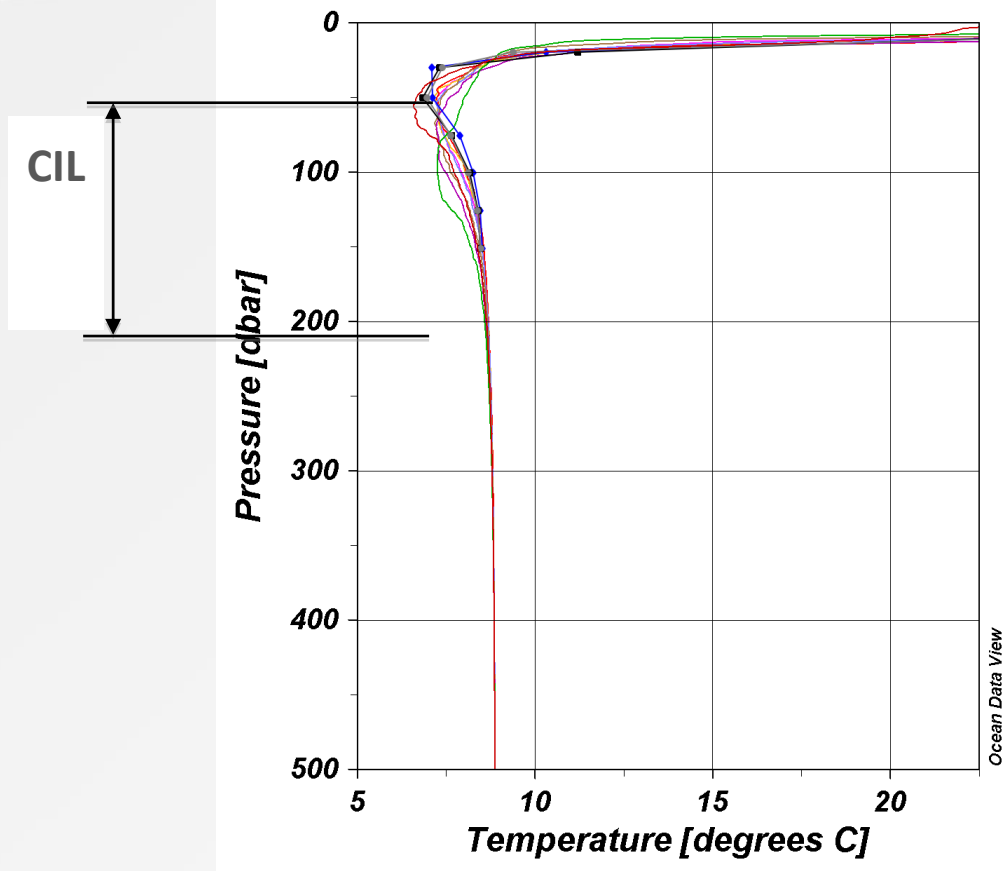


Figure 36 Typical Temperature profiles

Temperature profile (Figure 36) in Black Sea is characterized by presence of the Cold Intermediate Layer (CIL) – the layer where temperature goes below 8 C°. CIL position in depth and thickness is varying depending on location and time.

The TS diagrams in Figure 37 also indicate presence of CIL.

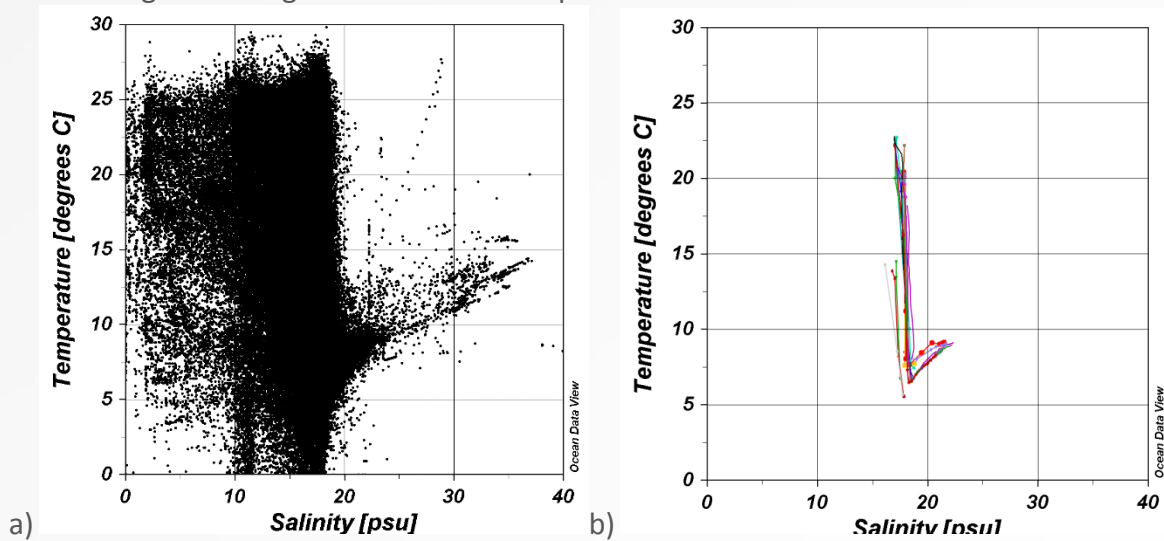


Figure 37 T-S diagram for entire dataset: a) scatter plot; b) typical shape of T-S curve

9.2.2. Data Quality assessment procedure for the Black Sea

The Black Sea dataset contains 20559 stations with not QC-ed Temperature and 18109 stations with not QC-ed Salinity. However presence of quality flags other than 0 does not guaranty that QC was applied in reality. There are many cases when data flagged as good appear to be wrong. Therefore Quality Assessment should be applied to the whole dataset.

Quality Assessment procedure for the Black Sea dataset included following steps:

- Identification of stations, which appear to be land as per ETOPO1 2' topography.
- Identification of stations without data. Such stations should be deleted.
- Identification of stations with Bottom Depth < 0.
- Identification of stations with profile depth > Bottom.
- Identification of stations with wrong date.
- Identification of stations with missing time.
- Range checks:
 - Depth < 0.
 - Temperature < 0 and QF<>4.
 - Temperature > 30 and QF<>4.
 - Salinity < 0 and QF<>4.
 - Salinity S > 39 and QF<>4.
 - Salinity > 23 out of Bosphorus area (28.8<Longitude<29.3, Latitude <41.6).
 - Temperature < 6 at depth > 200.
 - Temperature > 10 at depth > 200.
- Identification of duplicates.
- Flagging outliers
- Examination of profiles with density inversions

9.2.3. Quality assessment results

- Identified stations on land: 2379. Stations were excluded from further processing.
- Identified stations without data: 1474. Stations were deleted.
- Identified stations with Bottom Depth: 6. No action was applied - for attention of providers.
- Identified stations with profile depth > Bottom Depth: 18314. No action was applied - for attention of providers.
Note: in most of cases this missing Bottom Depth which is submitted as 0.
- Identified stations with wrong date: 49. Wrong dates were corrected.
In case of missing value the artificial date was assigned taking the date from the nearest station of the same cruise.
- Identified stations with missing time: 2362. For purposes of further data processing artificial time was assigned.
- Range checks:
 - Depth < 0: found at 26 stations.
The QF=4 was assigned to depth.

- Temperature < 0 and QF<>4:45 negative Temperature value meaning missing values (e.g. -999.99). Temperature value was deleted and respective QF set to 9.
- Temperature > 30 and QF<>4: 209.
Values =88 (meaning missing value) were deleted and respective QF to 9. Other values were flagged with QF=4.
- Salinity < 0 and QF<>4: 104.
Values -999.99 (meaning missing value) were deleted while the respective QF set to 9. For other negative values the QF=4 was assigned.
- Salinity S > 39 and QF<>4: 286.
QF=4 was assigned.
- Salinity > 23 out of Bosphorus area (28.8<Longitude<29.3, Latitude <41.6):286.
Data were flagged with QF=4.
- Temperature < 6 at depth > 200: 131.
Data were flagged with QF=4.
- Temperature > 10 at depth > 200: 256.
Data should be flagged with QF=4.
- Duplicates: 19064 duplicate groups were identified with the help of ODV. 17503 duplicates were already confirmed and properly chose stations marked for deletion. Note: the duplicates found with the help of ODV should be thoroughly examined before taking decision on deleting one of two. In most cases these are the same data submitted by different data providers, however one set can be QC-ed while other – not. The QC-ed data should retain in the collection. Sometimes duplicates found with ODV not appear to be real duplicates, for example in case kind of “time series” data when several measurements are performed at the same location at different time but in ODV collection time is missing or set to the same value by mistake.

Further data check procedure includes:

- Flagging of outliers which were not eliminated with range check
- Identification of profiles with inversions of density. Analyzing inversions and flagging respective parameter.

9.2.4. The Black Sea restricted data collection

Restricted data collection for the Black Sea initially contained 4370 stations that are 3.36% of total initial number of stations (125565 non-restricted + 4370 restricted). The collection contains 1287 stations without data, among which 1165 are duplicates. After excluding these stations the restricted data collection reduced to 3083 stations that are 2.95% of total stations accepted for further processing.

Black Sea restricted data	Number of stations	% to total
All stations	4370	3.4
Without duplicates and empty stations	3083	2.9

Tab. 7 Number of stations in the restricted Black sea data set.

First release of the aggregated data sets products – Friday 21 July 2017
sdn-userdesk@seadatanet.org – www.seadatanet.org

The restricted data collection covers period 1985–2010. Most of data are from the NE Black Sea, obtained in the last decade 2001–2010 (see Figure 38).

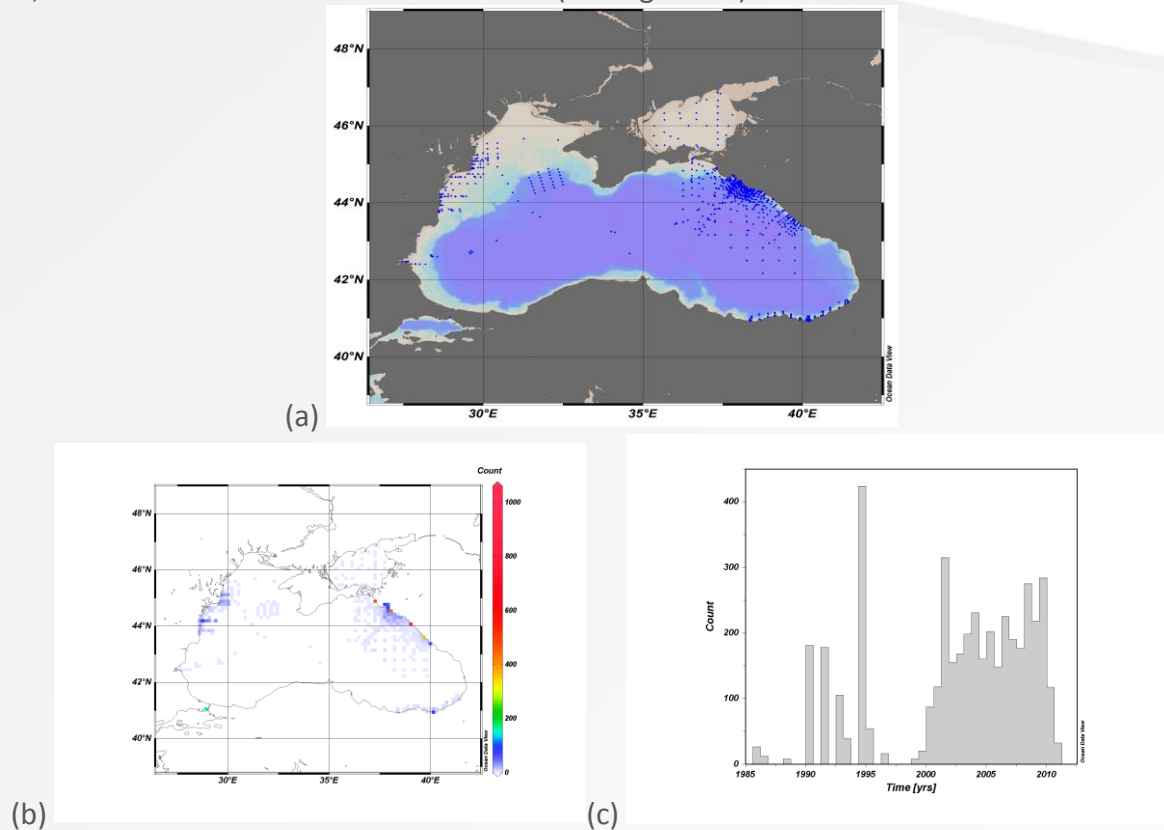


Figure 38 The Black Sea restricted data collection: (a) Data distribution map; (b) Data density map; (c) Annual distribution map.

Almost all data are within the ranges. Data on 1062 stations (~30%) are not QC-ed. 99% of the QC-ed data are flagged as good, however among them still wrong data were found. It means that QC should be (re)applied to all data regardless are they already flagged or not.

9.2.5. Conclusions

The initial Black Sea data set contains large number of not quality controlled data as well as erroneous data flagged as good. Also the dataset contains a lot of duplicates. In order to be utilized in oceanographic and ecological applications the dataset should be subjected to thorough quality assessment. The instructions on corrections of metadata and data and elimination of duplicates should be directed to data providers for applying as soon as possible otherwise next generations of the Black Sea dataset still will contain the same errors.

The Black Sea dataset potentially can be used for computing of climatologies, however it is necessary to take into account that availability of data drastically decreases with depth (see Figure 39).

The density of data coverage for selected IODE depth levels is presented at Figure 39. In principle data are scattered more or less evenly within the respective bathymetry contours,

however the density drastically decreases with depth, which may result in low quality of climatologies for deep levels. Another problem is variability of CIL (see above) which may lead to increasing of error for climatologies in upper layer despite rather good data coverage.

Considering the data coverage it is planned to calculate seasonal climatic maps of Temperature and Salinity for depth levels 0, 10, 20, 30, 50, 75, 100, 150, 200, 250 m and annual maps for depth levels 300,400, 500, 600, 800, 1000, 1200, 1500, 2000 m.

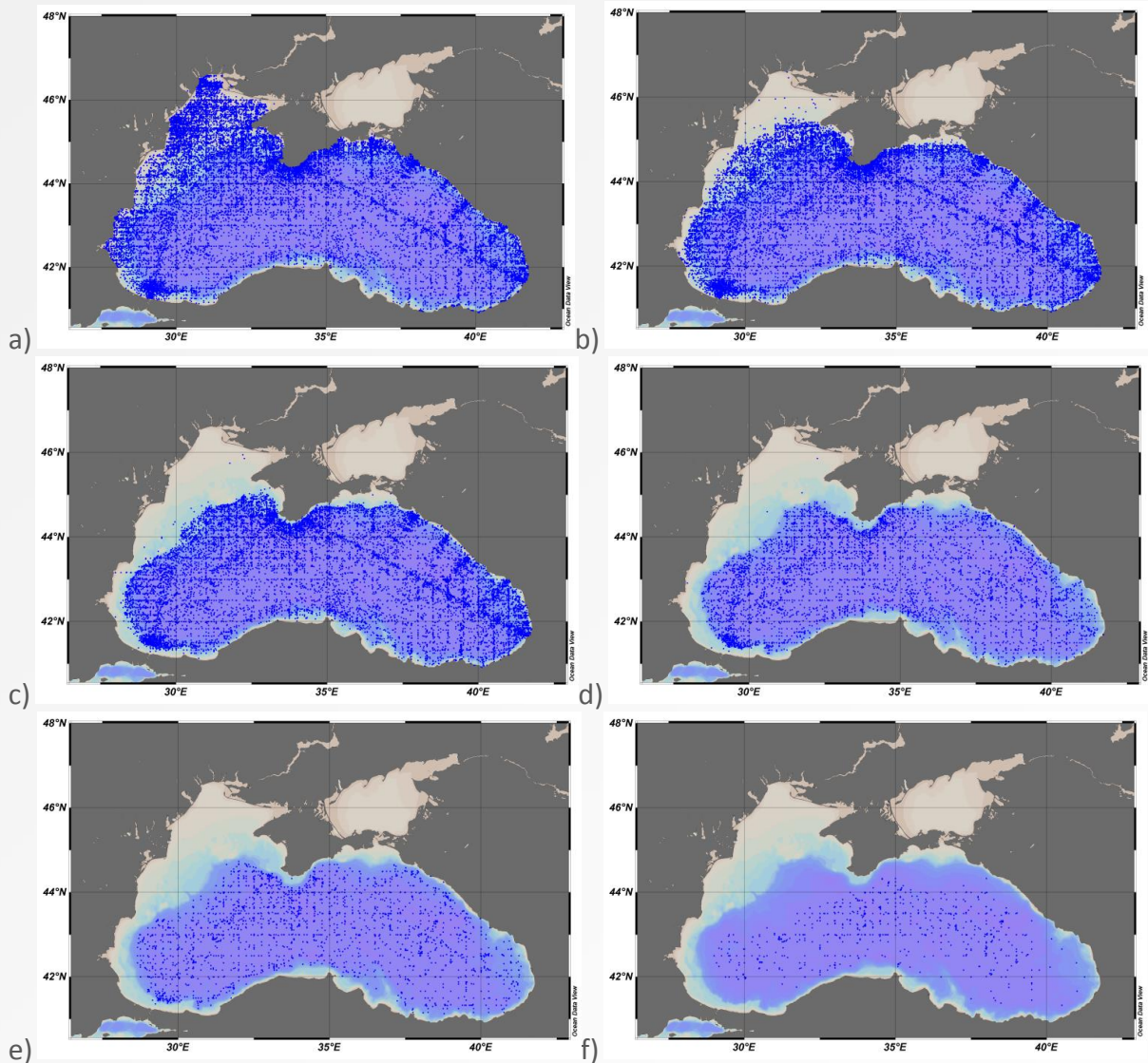


Figure 39 Maps of data coverage for different IODE depth levels: a) 20m; b) 50m; c) 100m; d) 500m; e) 1000m; f) 2000m.

Contribution of the current restricted data collection to the Black Sea climatologies appears to be not very significant; however it may increase in future in case more data providers will allow access to their restricted data.

9.3. Arctic Sea

The procedure for the quality assessment analysis is built according to some criteria.

- Outliers: data outside of the regional range defined for temperature and salinity parameters are excluded (QF 4).
- Density inversion, when temperature and salinity measurements are available.
- A visual QF control is applied on all the dataset, by time period, by latitude and longitude bands to focus on specific area. For all the spikes, density inversions, gradients, doubtful data, that are detected, the QC is changed to 3 or 4.
- Data on land were also identified to be listed.

A list of all those anomalies has to be sent to the NODCs for correcting the QFs.

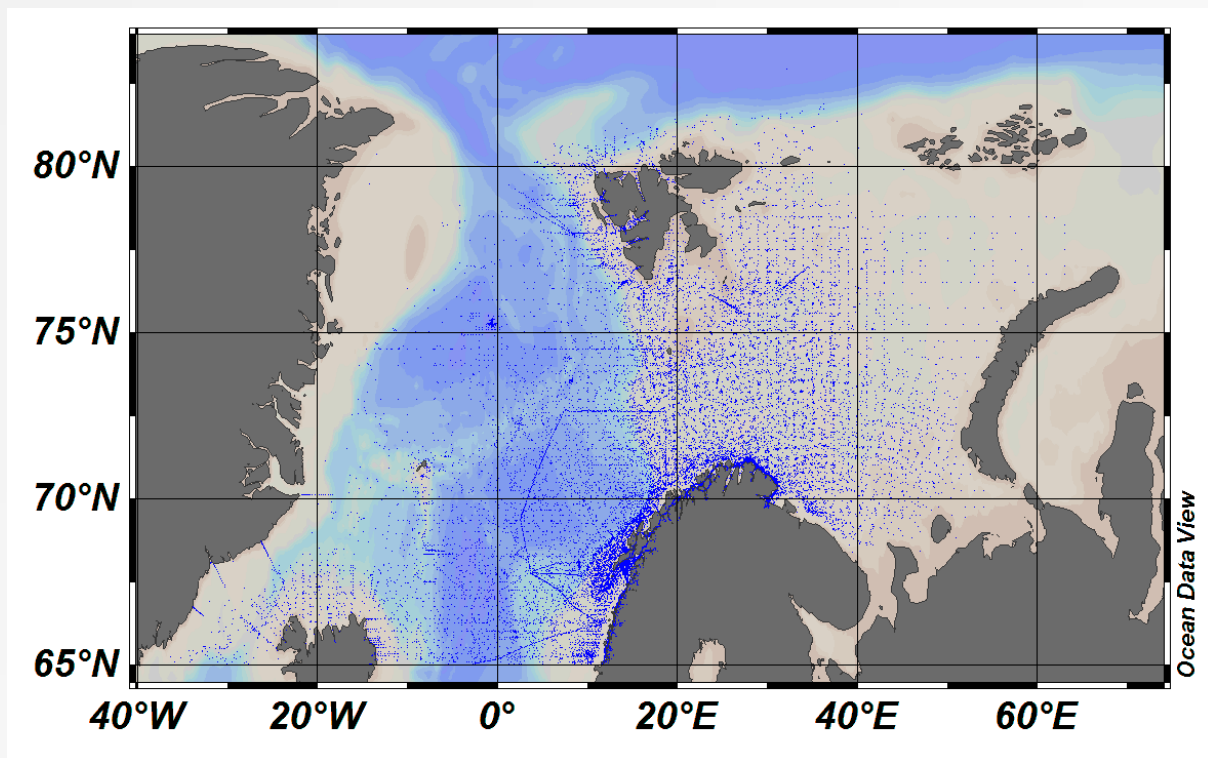


Figure 40 Arctic1900-2013 TS data distribution map V1.1 aggregation for the time period 1900-2013

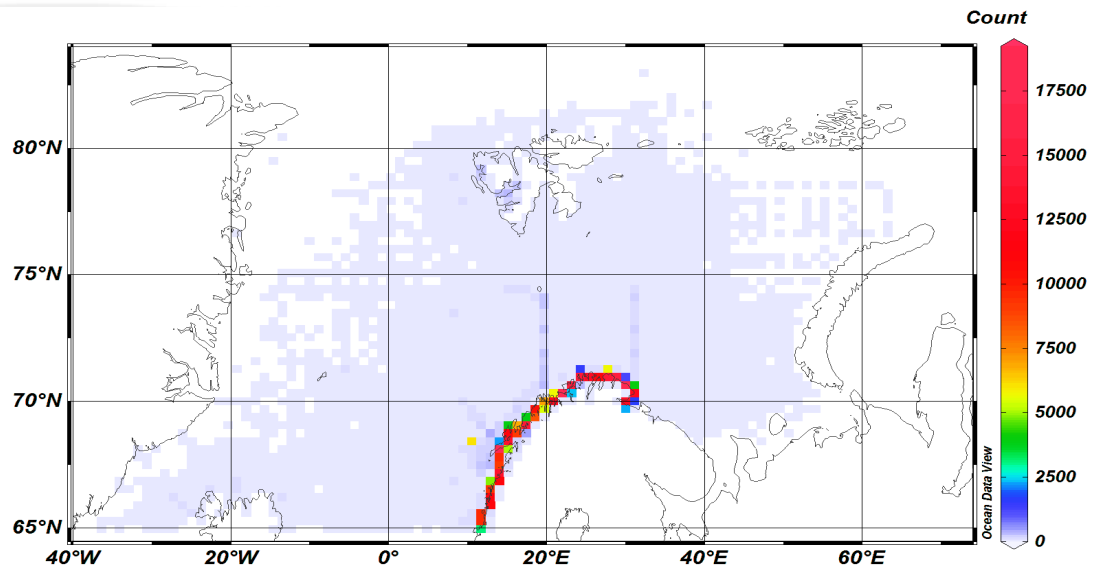


Figure 41 Arctic1900-2013 TS data density map V1.1 aggregation for the time period 1900-2013

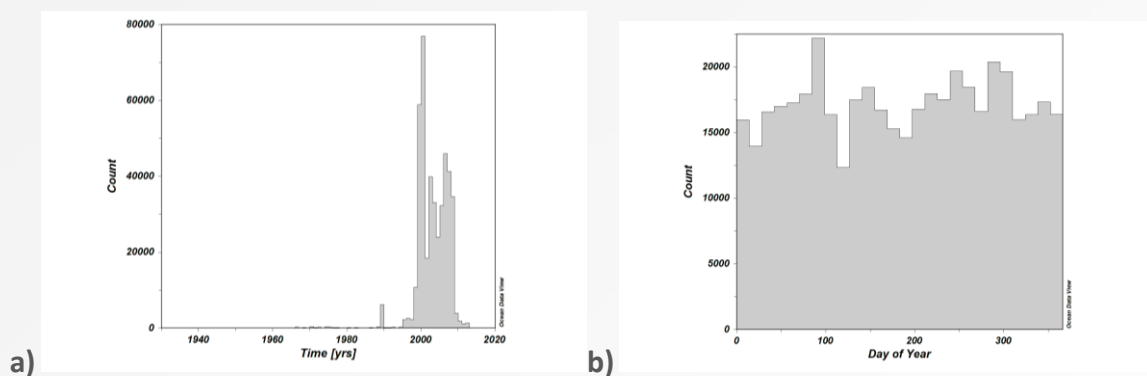


Figure 42 Arctic1900-2013 (a) Annual data distribution and (b) seasonal data distribution for the time period 1900-2013

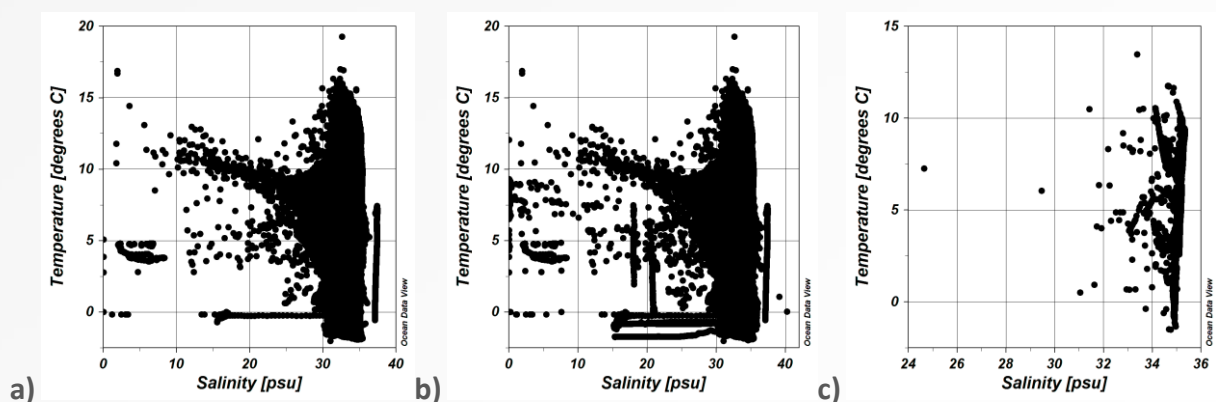


Figure 43 Arctic1900-2013 TS data collection for the time period 1900-2013: (a) TS diagram after range check analysis; (b) TS diagram considering only data with QC flags = 1 (good) and 2 (probably good) for depth, T and S; (c) TS diagram considering only data with QC flags = 0 (no quality control) for depth, T and S.

9.4. Baltic Sea

9.4.1. General Characteristics of the Baltic Sea historical data set

The Baltic Sea historical data set contains just over 209000 CDIs with almost 12000000 values for both salinity and temperature. There are also about 8600 restricted CDIs with a total of about 65000 salinity and temperature values. This means almost all this data is free, less than 0.5% of the data is restricted. Most of the data are from profiles, dots in Figure 44(a), but there are also some data that are from trajectories (ferry box system), solid lines in Figure 44(a).

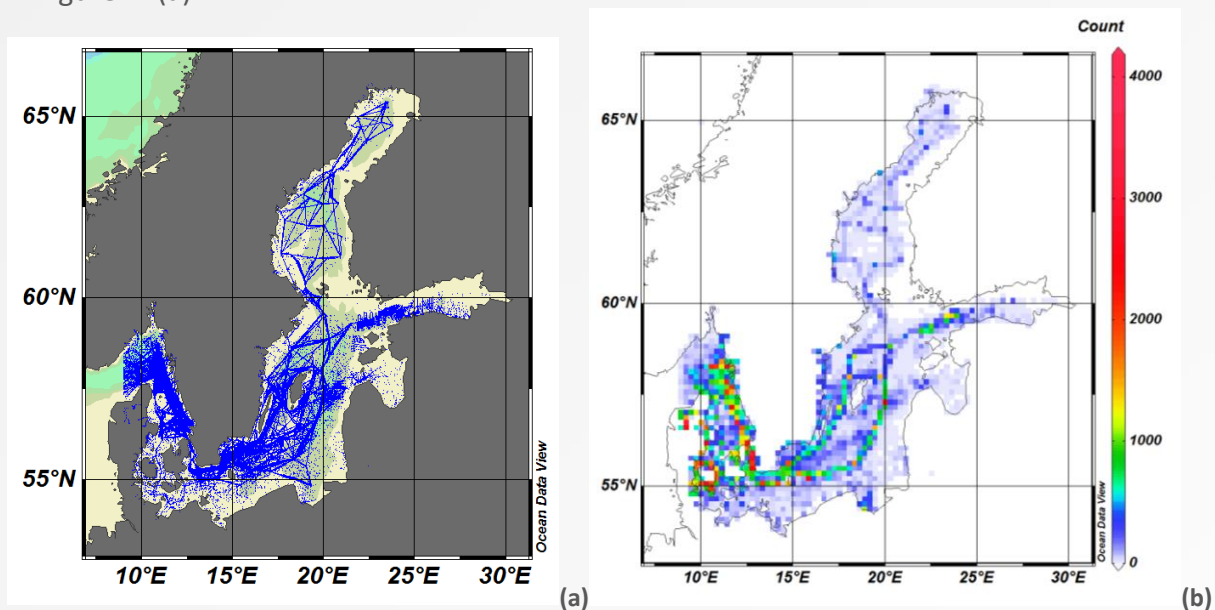


Figure 44 TS data collection for the Baltic Sea in the time period 1990-2013: (a) Data distribution map; (b) Data density map.

Data distribution map (Figure 44a) show a good geographical spread but with a few coastal areas with no or almost no data. Data density map (Figure 44b) is heavily dominated by trajectory data (ferry box system), due to the large number of data points in this type of data. Data density map does not show the vertical data density, this means the trajectory data can create a false illusion of lots of data, when in reality the trajectory only contains data from one depth and can have a small time coverage as well.

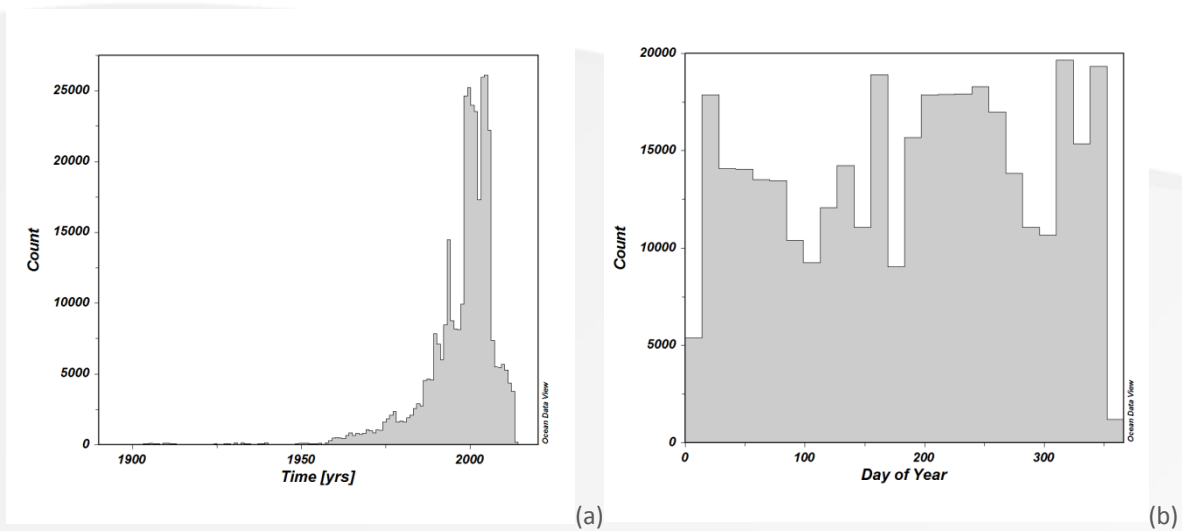
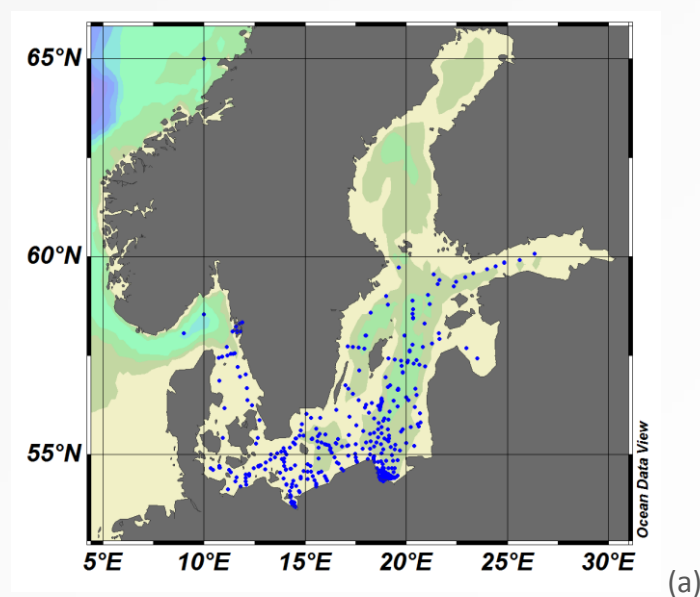


Figure 45 (a) Annual data distribution and (b) seasonal data distribution for the time period 1990-2013 in The Baltic Sea.

Annual data distribution (Figure 45a) shows that there are few measurements up until about 1960. In the 1980s and 1990s the data points increases and stays on a relatively high level up until the latest years where there is a natural time lag between sampling and until data becomes available in the SeaDataNet system. The spikes that can be seen in the late 1990s and early 2000s are from ferry box trajectory data, containing large amounts of data points. Seasonal data distribution (Figure 45b) shows an even spread during the year.

The restricted data comes from the Baltic Proper and also a few from Skagerrak and Kattegat, Figure 46(a). The temporal distribution (c) shows that almost all of the restricted data comes from the period 1960-1990. Hopefully it will be possible to release most of this data as unrestricted in the near future, considering that the data is quite old and is expected to be easier to make free to use compared to more recent data.



First release of the aggregated data sets products – Friday 21 July 2017
sdn-userdesk@seadatanet.org – www.seadatanet.org

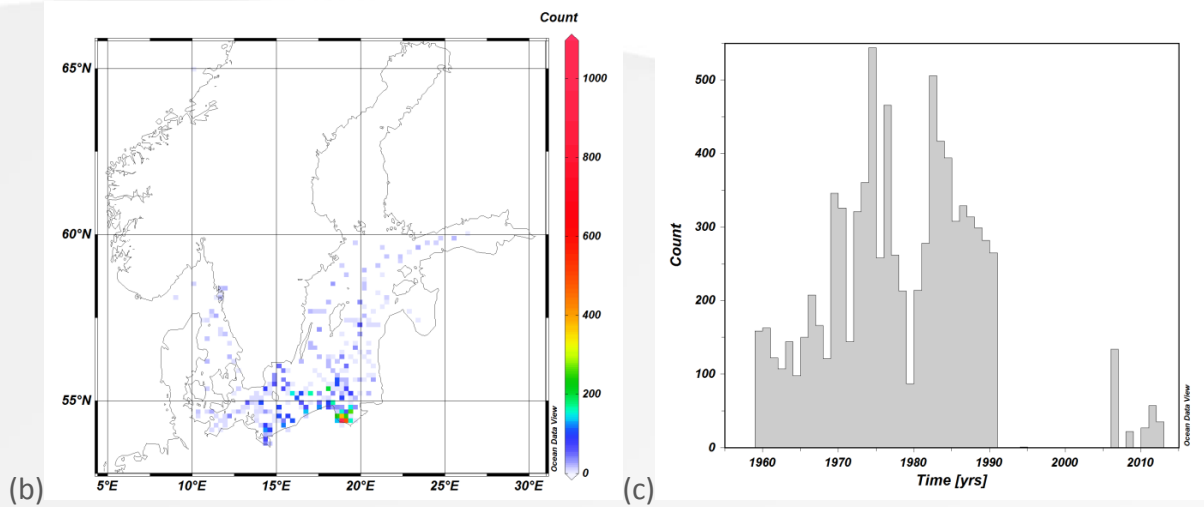


Figure 46 The Baltic Sea restricted data collection. (a) Data distribution map, (b) data density map and (c) annual distribution map.

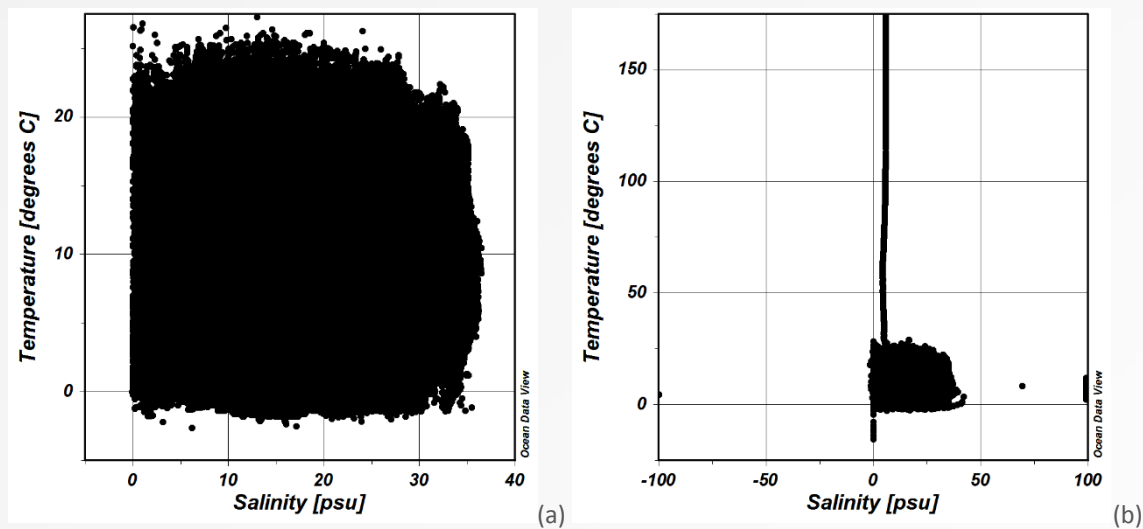


Figure 47 Baltic Sea TS data collection for the time period 1990-2013: (a) TS diagram after QC; (b) TS diagram before QC.

Temperature-Salinity scatter plot before quality control, Figure 47(b), show that are some obvious outliers that are easy to detect and remove. TS scatter plot after quality control show that no obvious outliers are present. It also shows the large range in both salinity, 0 - 36, and temperature, -2°C – 26°C. These large variations make quality control harder, and it makes range checks almost useless. This can be handled by splitting the data into subsets, either by choosing a subset in time or a sub-region with less variation than the whole data set, or both combined.

9.4.2. Data Quality assessment procedure for The Baltic Sea

Salinity has a large geographical variation, from down to 0 in the north up to 36 in the southwest, Figure 48(a). To handle this, and make QC more manageable, salinity data was divided into sub-regions, Figure 48(b). For each region all data were then filtered for one year at a time and profiles were visually inspected in ODV to discover spikes, outliers and unstable profiles. We also calculated and plotted density to easier find the unstable profiles. For temperature even smaller sub-regions were used to keep the number of profiles in each region lower. Data were then filtered out for one or two months at a time (all years) to handle the large seasonal variation; below 0°C in the surface during winter, and up to above 25°C during summer. Profiles were then visually inspected in ODV to find spikes and outliers. We did the same procedure for all data, not considering a difference between quality flags 0 and 1; since it is well known that quality controlled data still can contain errors. Obvious bad data were flagged with quality flag 4 (bad), and suspicious data were flagged with flag 3 (probably bad).

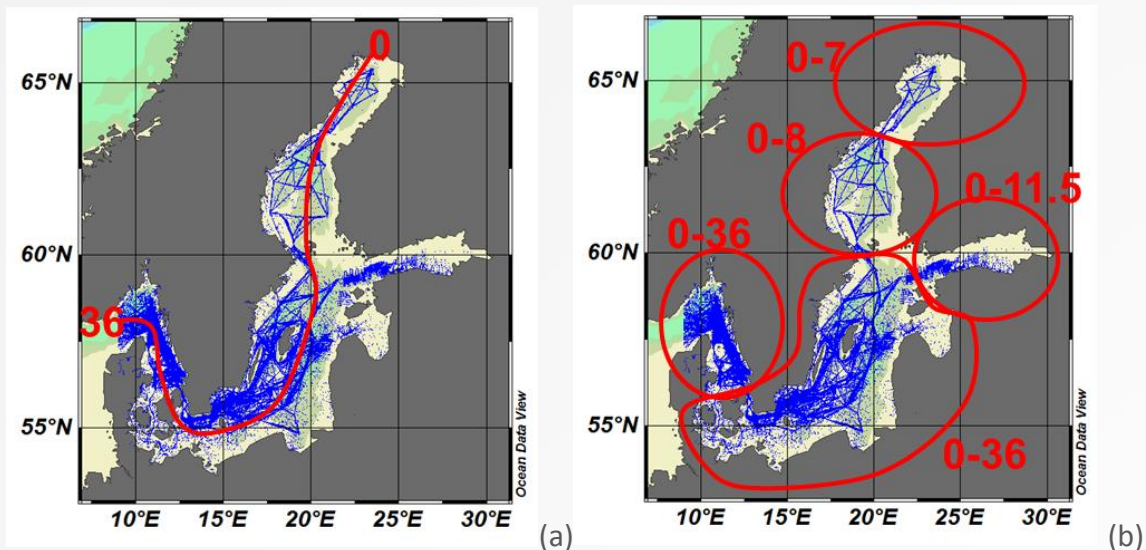


Figure 48 (a) Salinity variation in The Baltic Sea, (b) sub-regions used to easier handle the quality control.

A few other more obvious errors were found; 1-2 stations with wrong position, for example on land. Some profiles were too deep and a few had negative depths. In total about 1600 CDIs contained bad/suspicious data, which is less than 1% of the total number of CDIs. For the restricted data 18 CDIs contained suspicious data, about 0.2% of the total number of restricted CDIS.

9.4.3. Conclusions

The quality of the data set from The Baltic Sea is high. Less than 1% of all data were flagged as suspicious after a thorough quality control. Most of the data are free to use, only a small portion are restricted.

Data is well distributed both spatial and temporal which should work for monthly climatologies in DIVA.

Some early tests in DIVA have been made with the cleaned data set and the preliminary results look promising. However there are some signs that the high density data from the ferry box data impact the result in DIVA a bit too much. Since this type of data includes so many data points but mostly from just a short time interval, it is likely that these measurements have a too big impact on the DIVA end result than what is desired. This issue needs to be discussed and dealt with. Possible solutions are to remove this data type completely from the DIVA runs, or to try to reduce the data density by for example just keeping every third data point from this data type.

9.5. North Sea

9.5.1. General characteristics of the North Sea Historical data collection

The North Sea historical data set contains 751844 CDIs pointing to data freely available (hereafter referred to as the “historical dataset”). The restricted dataset contains 830513 CDIs (hereafter the “restricted dataset”). These two numbers being of the same order of magnitude, it is worth having a look in Figure 49 to the distribution of the data in parallel.

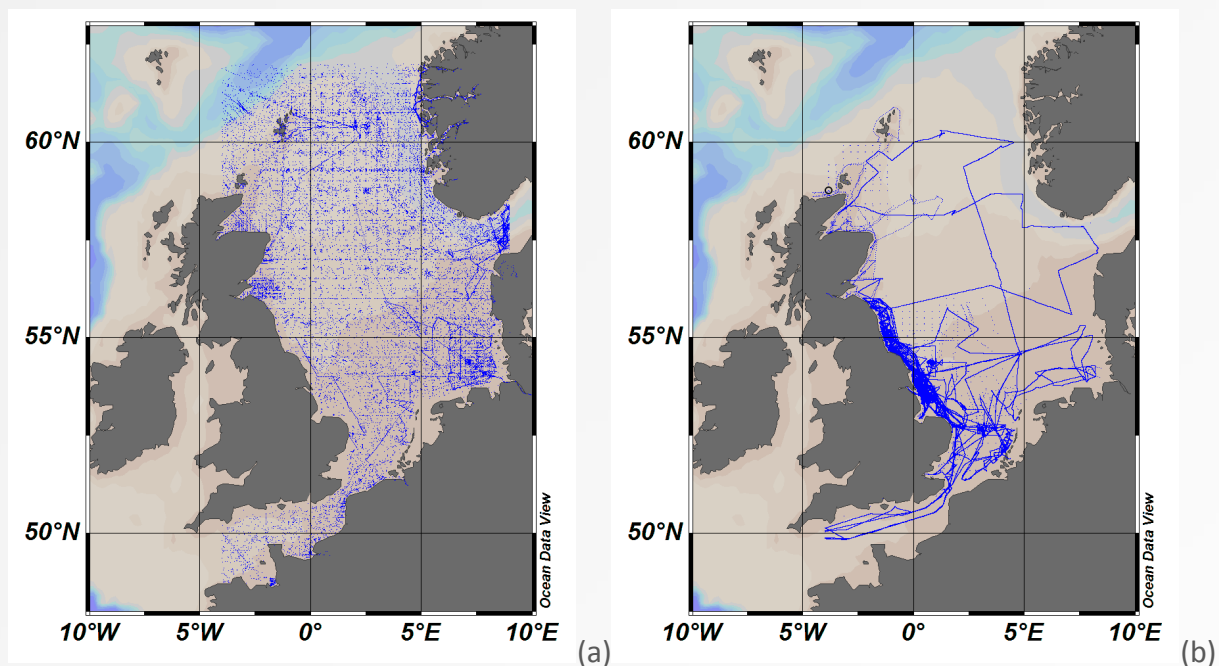


Figure 49 Location of stations in the TS data collections for the North Sea for the period 1900-2013: (a) freely accessible data (“historical dataset”), (b) restricted dataset.

The geographical coverage of stations in the historical dataset is almost uniform, with the highest number of stations in the Forth Estuary, the Moray Firth, the German Bight, the entrance of the Skagerrak and in Norwegian coastal waters (incl. fjords). Except for a few transects in the North Sea and especially in its Southern Bight, the restricted data are located along the coast of Great-Britain, from the Norfolk up to Edinburgh.

The density of information (see Figure 50) gives a slightly different picture for the historical dataset than the map showing the location of stations (Figure 49).

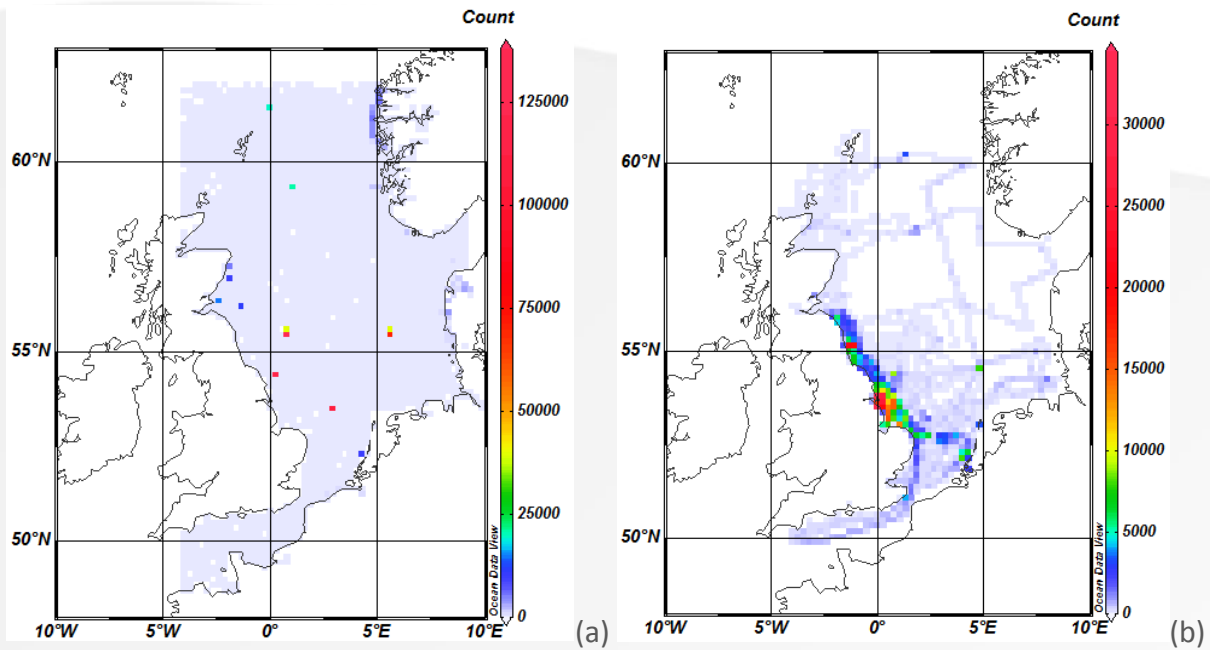


Figure 50 Data density maps of the TS data collections for the North Sea for the period 1900-2013: (a) historical dataset (b) restricted dataset.

The data density map of the historical dataset shows a few stations with a huge amount of data. These stations were part of an intense measurement effort at the end of the 1988 and in 1989, as shown in Figure 51 and Figure 52.

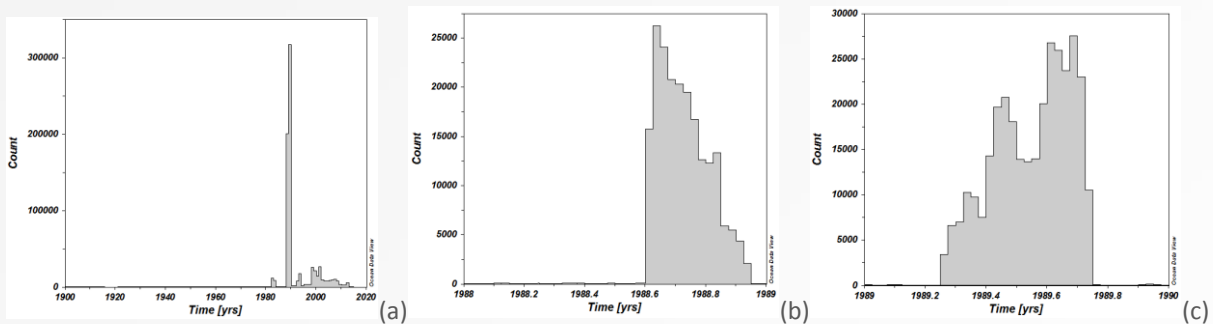


Figure 51 Time histogram of the data in the historical dataset: (a) full dataset, (b) focus on 1988 (200343 CDIs), (c) focus on 1989 (317290 CDIs).

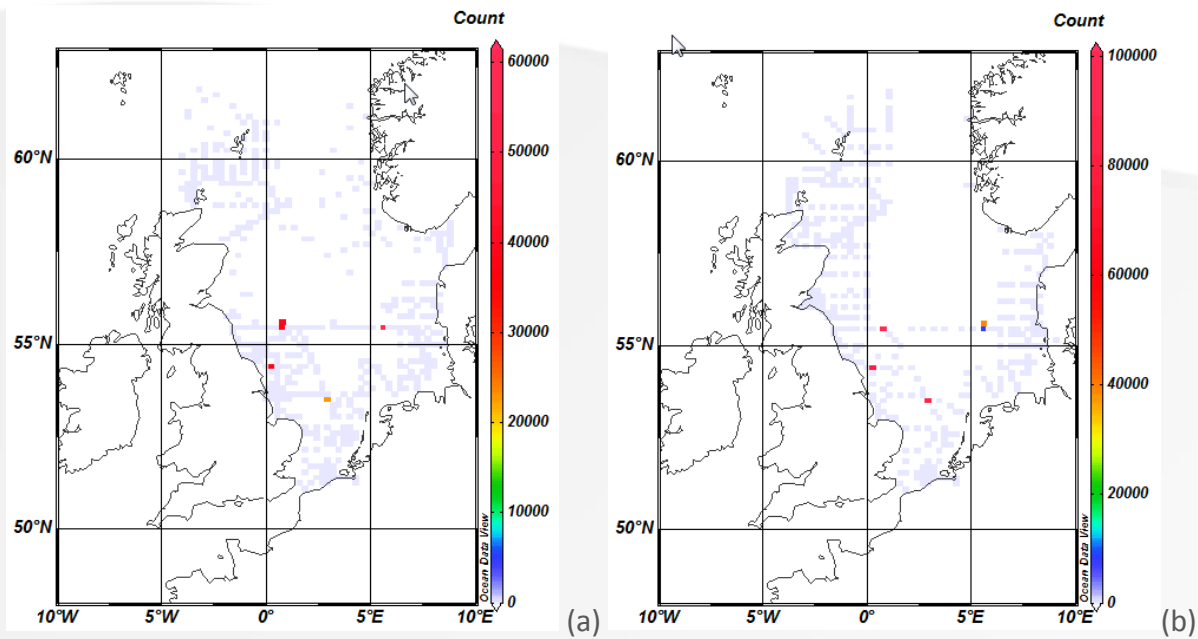


Figure 52 Data density maps of the historical dataset: (a) focus on 1988 (200343 data points), (b) focus on 1989 (317290 data points).

When excluding 1988 and 1989 data from the historical dataset, the data density map in Figure 53a, still shows a few “hot spots” but the more intense measurement effort along the Danish and Norwegian coasts appear clearly.

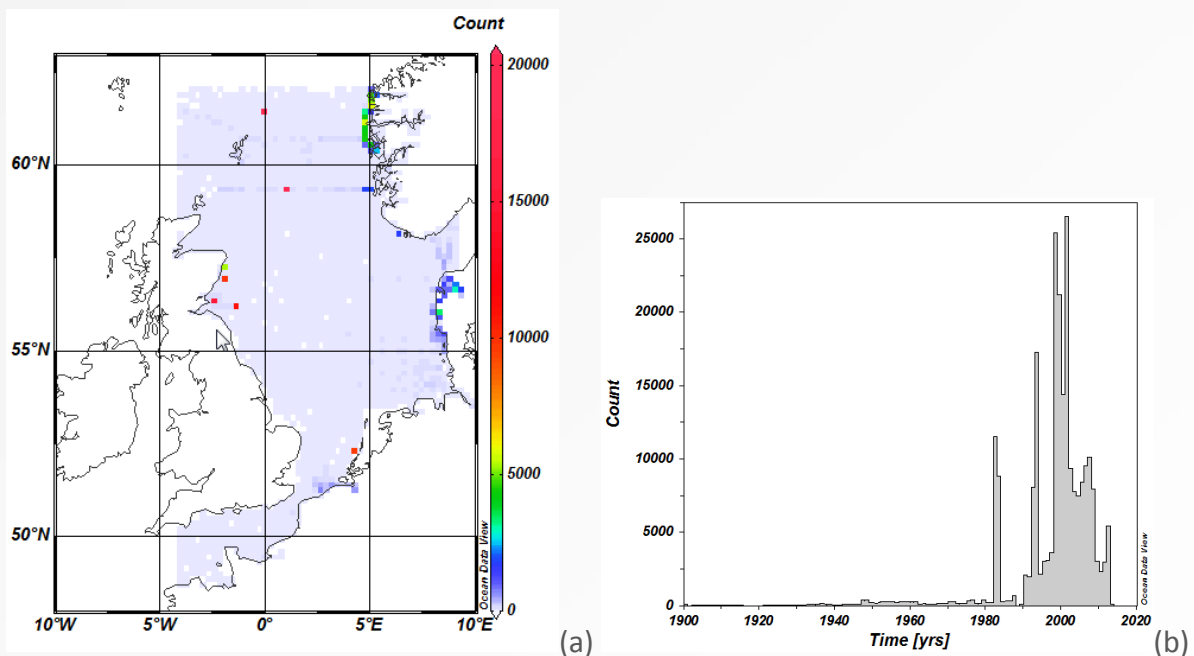


Figure 53(a) Data density maps of the historical dataset, from 1900 till 2013, excluding 1988 and 1989; (b) distribution over time of the data in the historical dataset, from 1900 till 2013, excluding 1988 and 1989.

Similarly, the data in the restricted dataset mainly correspond to two periods of intense measurement: 1988–1990 (238675) and 1993–1995 (513446) as displayed in Figure 54 and Figure 55 relatively.

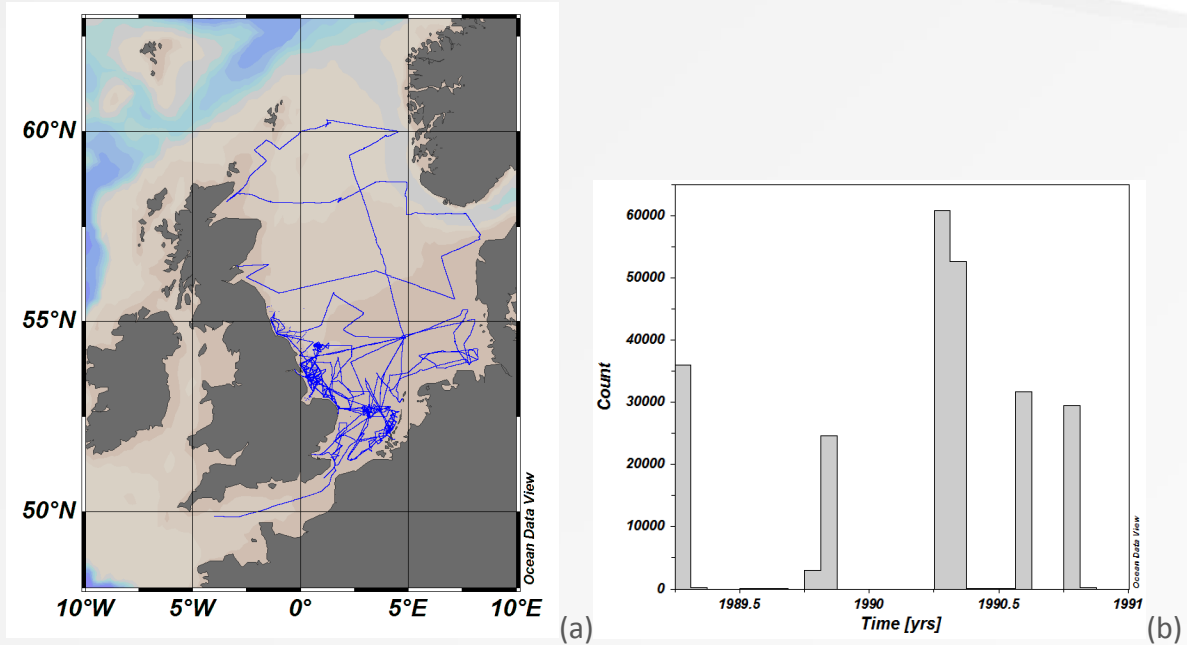


Figure 54 (a) Location of the stations in the restricted dataset for the period 1989–1990, (b) Distribution of data over time in the restricted dataset for the period 1989–1990.

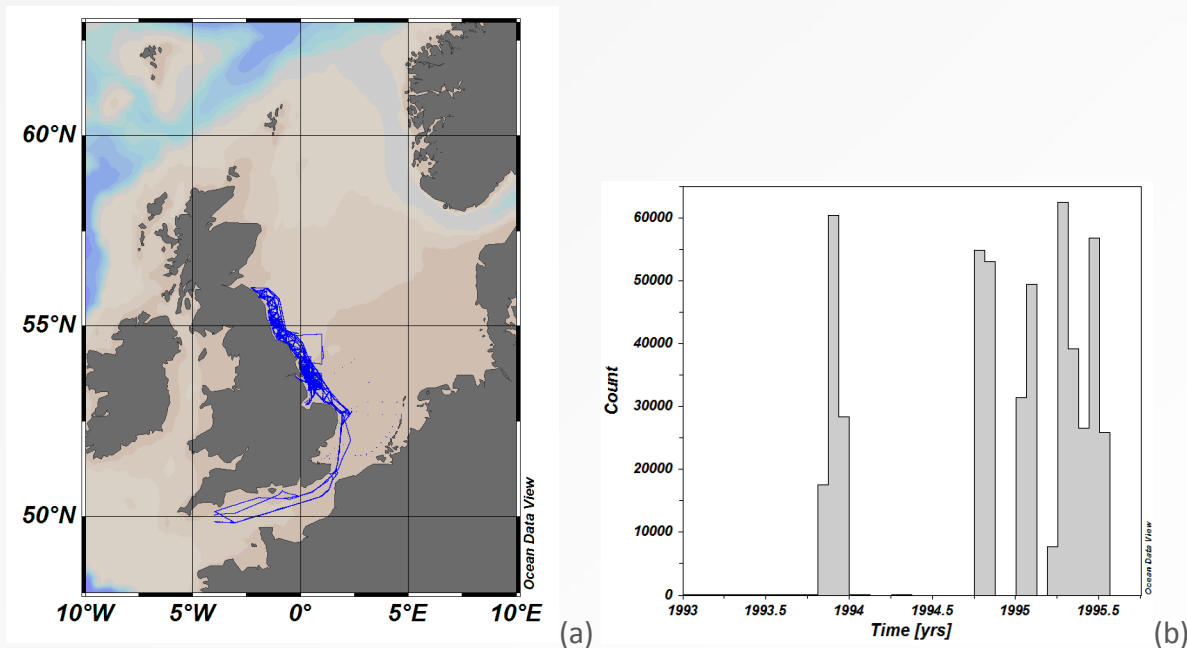


Figure 55 (a) Location of the stations in the restricted dataset for the period 1993–1995, (b) Distribution of data over time in the restricted dataset for the period 1993–1995.

Finally, the restricted dataset shows in seasonal data distributions of Figure 56b an emphasis on the Spring and Winter seasons while the historical dataset contains a larger occurrence of measurements during Fall (Figure 56a).

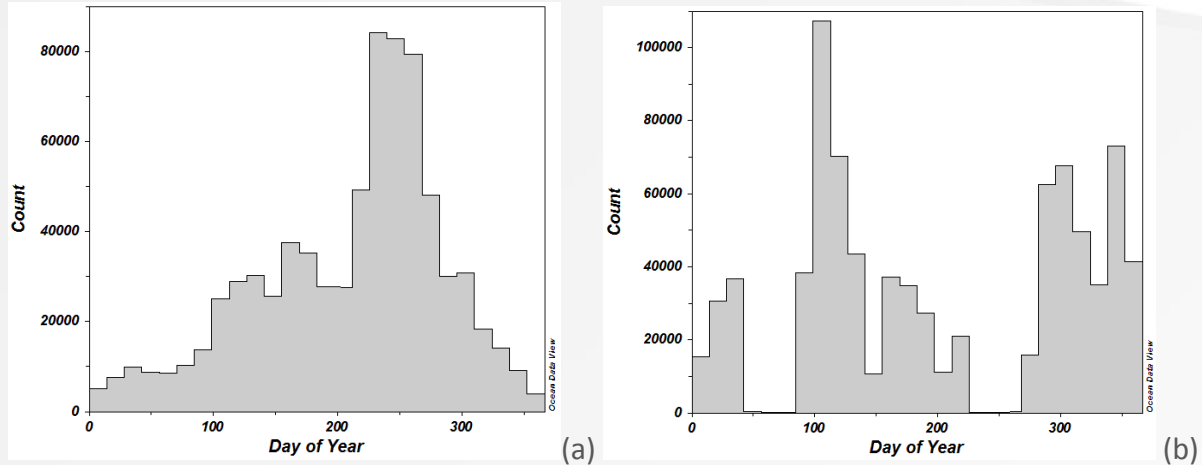


Figure 56 Seasonal distribution of the data: (a) historical dataset, (b) restricted dataset.

This comparison of the spatial and time distributions of the data in both the historical and the restricted datasets shows that both datasets exhibits peculiarities (high density of measurements either at specific locations or during a limited period of time) that should be taken with care in any further use of these data.

9.5.2. Quality of the historical dataset

The historical dataset contains 5,422,771 TS-pairs of which 5,385,855 (99.32%) are QC-flagged “1” (good) or “2” (probably good).

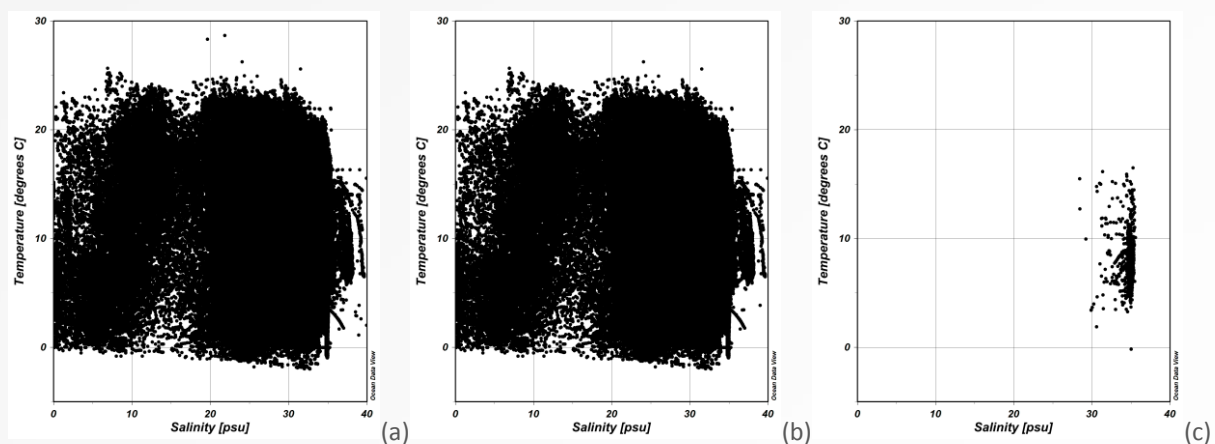


Figure 57 North Sea TS historical data collection: (a) TS diagram after range check analysis; (b) TS diagram considering only data with QC flags = 1 (good) and 2 (probably good) for T and S; (c) TS diagram considering only data with QC flags = 0 (no quality control) for T and S.

The dataset contains much more temperature data than salinity data as summarized by Tab. 8. Although Figure 57 show that the NODCs apply conscientiously and consistently quality control procedures. Some data and profile escape the screening, as evidently shown in Figure 58

	TOT	QF0	QF1	QF2	QF≥3
T	12,497,681	2,268 (0.02%)	11,739,794 (93.94%)	4 (-)	755,615 (6.04%)
S	5,489,802	16,584 (0.30%)	5,451,843 (99.31%)	69 (-)	21,306 (0.39%)

Tab. 8 Number of Temperature and Salinity data points for the North Sea historical data collection and their subdivision according to Quality Flags (QF) 0, 1, 2 and from 3 to 9.

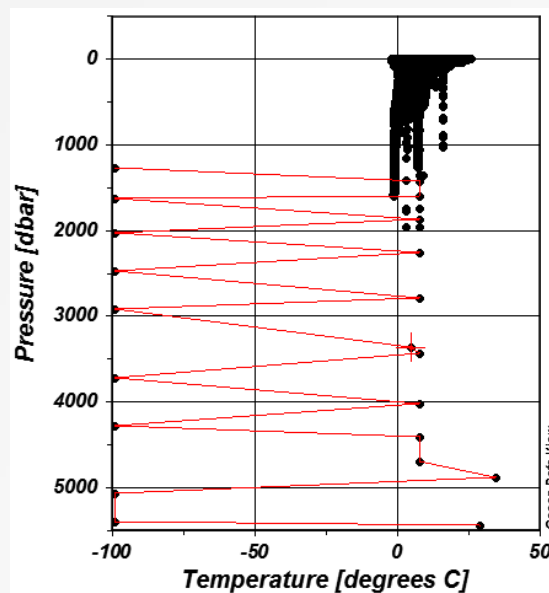


Figure 58 Temperature data from the North Sea historical data collection where original QF=1, exhibiting data that obviously weren't correctly flagged.

A rough application of strict QC checks results in about 6% of the data originally flagged as good to be rejected (QC flag 3, “probably bad”, or 4, “bad”). These anomalies will be reported to the corresponding collating centres. It is expected that the actual number of data erroneously flagged as good will be lowered after assessment with the collating centres.

9.5.3. Quality of the restricted dataset

The restricted dataset contains 898,030 TS-pairs of which 850,123 (94.67%) are QC-flagged “1” (good) or “2” (probably good). Figure 59 shows TS diagrams of the entire restricted dataset and of data with QF=1&2 and QF=0.

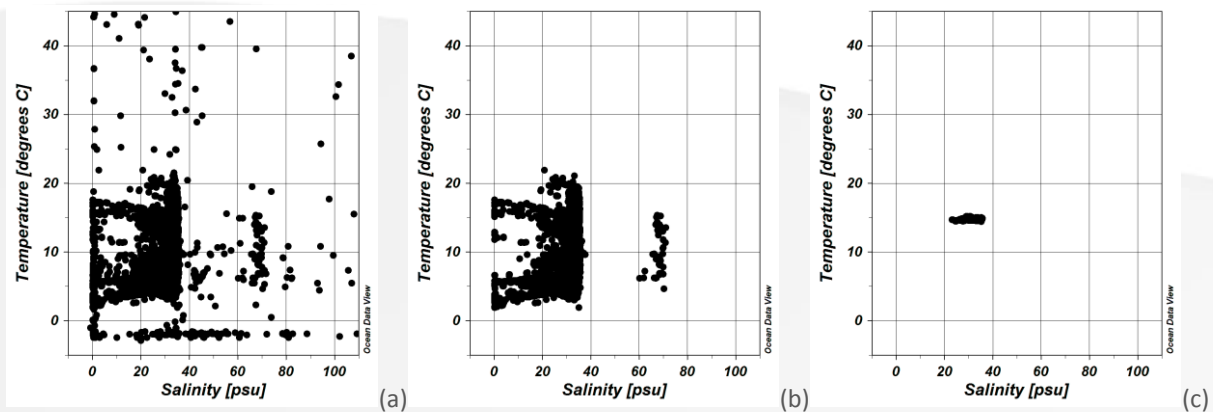


Figure 59 North Sea TS restricted data collection: (a) TS diagram of the full dataset; (b) TS diagram considering only data with QC flags = 1 (good) and 2 (probably good) for T and S; (c) TS diagram considering only data with QC flags = 0 (no quality control) for T and S.

	TOT	QF0	QF1	QF2	QF≥3
T	917,001	1,605 (0.18%)	879,250 (95.88%)	0 (-)	36,146 (3.94%)
S	923,418	21,679 (2.35%)	862,328 (93.38%)	0 (-)	39,411 (4.27%)

Tab. 9 Number of Temperature and Salinity data points for the North Sea restricted data collection and their subdivision according to Quality Flags (QF) 0, 1, 2 and from 3 to 9.

A detailed investigation of the bad data erroneously flagged as good has shown that it concerns less than 0.01% of the data.

9.5.4. Conclusions

The quality of the North historical dataset is rather good. Data from two collating centres show a too high occurrence of bad data erroneously flagged as good and a cross check will be performed with the these centres.

There is a significant number of data with restricted access rules: 13.52% of the total TS-data. The overall quality of the restricted dataset is very good.

The spatial and time distribution of both datasets present peculiarities that should be kept in mind when using the datasets.

9.6. The North Atlantic Ocean

9.6.1. General Characteristics of the North Atlantic Ocean historical data set

The North Atlantic Ocean historical data set contains just over 1049547 CDIs. There are also about 28338 restricted CDIs that mean less than 3% of the whole dataset is restricted. Most of the data are from profiles, dots in Figure 60(a) and are distributed on the east part of the North Atlantic Ocean (Figure 60 (b)).

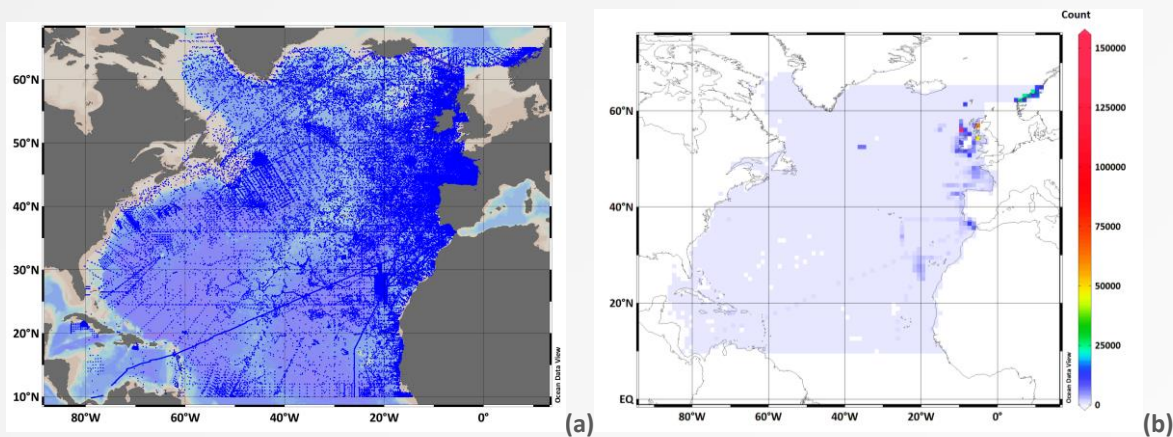


Figure 60 TS data collection for The North Atlantic Ocean in the time period 1900-2013: (a) Data distribution map; (b) Data density map.

Data distribution map show a good geographical spread with a best coverage on the east part, mainly close to the coastal areas and in the Bay of Biscay (Figure 60 (b)).

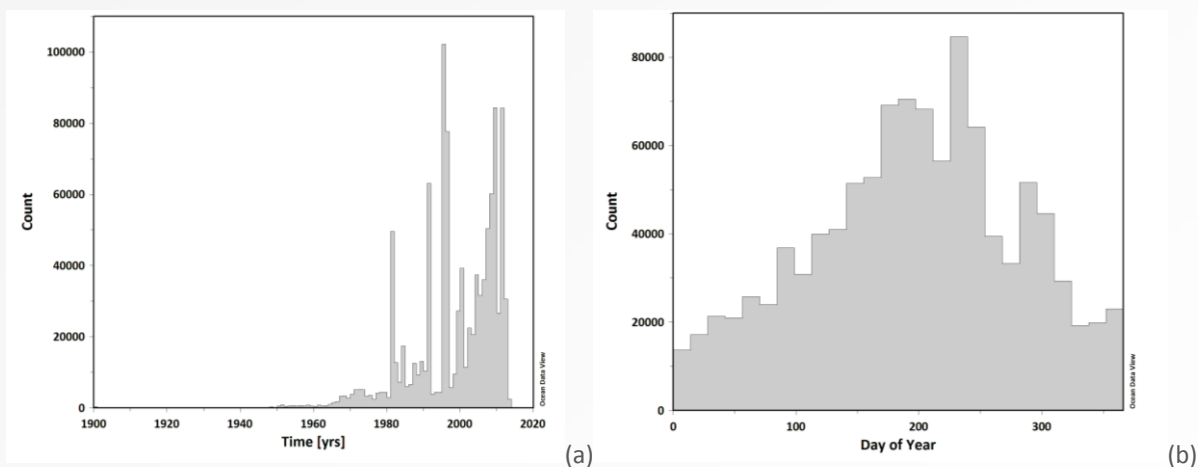


Figure 61 (a) Annual data distribution and (b) seasonal data distribution for the time period 1900-2013 in The North Atlantic Ocean.

Annual data distribution Figure 61 (a) shows that there are few measurements up until about 1960. Peaks can be observed in 1996-1997 and some new recent data in 2010-2012, which is good. Concerning the seasonal distribution Figure 61 (b), the peak of the dataset is observed during the summer time.

Due to the large number of data, the North Atlantic area has to be divided on 3 time periods to display the data plots: 1900-1999 (Figure 62), 2000-2008 (Figure 63), and 2009-2013 (Figure 64).

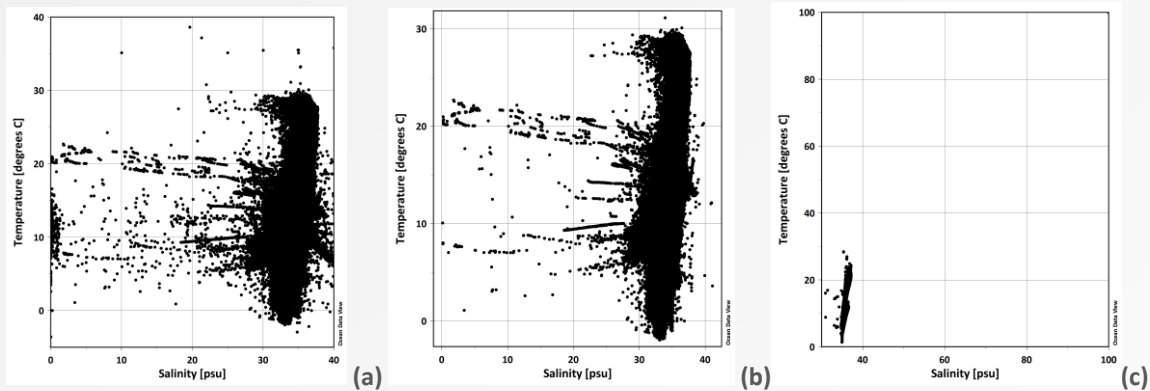


Figure 62 North Atlantic TS data collection for the time period 1900-1999: (a) TS diagram after range check analysis; (b) TS diagram considering only data with QC flags = 1 (good) and 2 (probably good) for T and S; (c) TS diagram considering only data with QC flags = 0 (no quality control) for T and S.

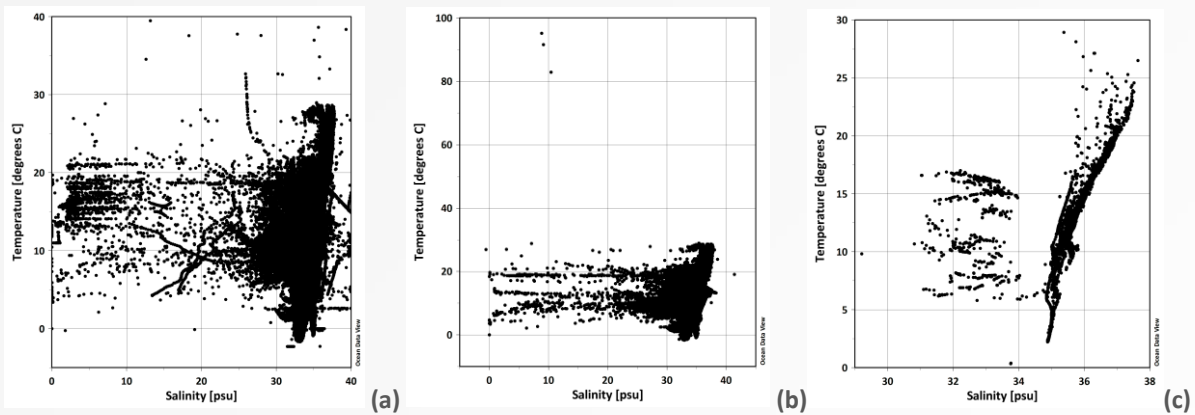


Figure 63 North Atlantic TS data collection for the time period 2000-2008: (a) TS diagram after range check analysis; (b) TS diagram considering only data with QC flags = 1 (good) and 2 (probably good) for T and S; (c) TS diagram considering only data with QC flags = 0 (no quality control) for T and S.

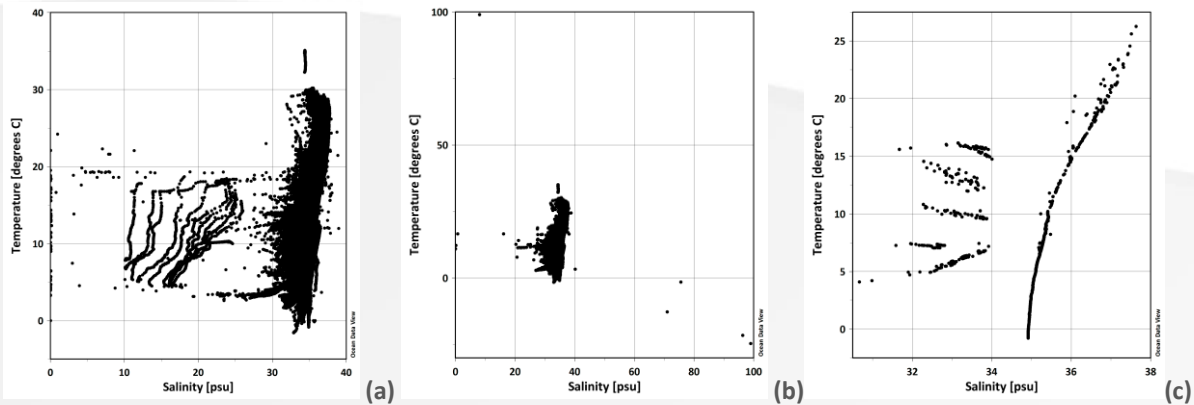


Figure 64 North Atlantic TS data collection for the time period 2009-2013: (a) TS diagram after range check analysis; (b) TS diagram considering only data with QC flags = 1 (good) and 2 (probably good) for T and S; (c) TS diagram considering only data with QC flags = 0 (no quality control) for T and S.

Most of the measurements present a good quality. Nevertheless, some QF0 (Figure 62, Figure 63c, Figure 64c) have been observed and some bad measurements have still good QF (1). The majority of temperature and salinity measurements (99.2%) have QF1, less than 0.5% of measurements are considered as doubtful or bad by the NODCs (see measurements statistics about quality flags in Tab. 10).

	TOT	QF0	QF1	QF2	QF3-9
T	46596335	145511 0,31%	46231467 99,21%	781 0,001%	218546 0,47%
S	16121430	718517 4.5%	15360727 95.3%	0	42186 0.3%

Tab. 10 North Atlantic TS measurements collection for the time period 1900-2013: number of measurements (and percent) for temperature and salinity sorted by quality flag.

Temperature-Salinity scatter plots (for QF 1&2) before quality control (Figure 62b, Figure 63b, Figure 64b) show that are some obvious outliers those are easy to detect and remove. The large variability of both salinity and temperature makes the quality control difficult, thus the data set has been split into sub-sets, either in time or in space (sub-regions) or both combined, with a smaller variation than the whole dataset.

Those bad data clearly appear in the plots but they represent (after applying QC procedure) less than 0.3 % of the good data local_cdi_ids (9 Edmo_codes are concerned by those anomalies). Figure 66 show temperature and salinity diagrams taking into account some corrections on the QC.

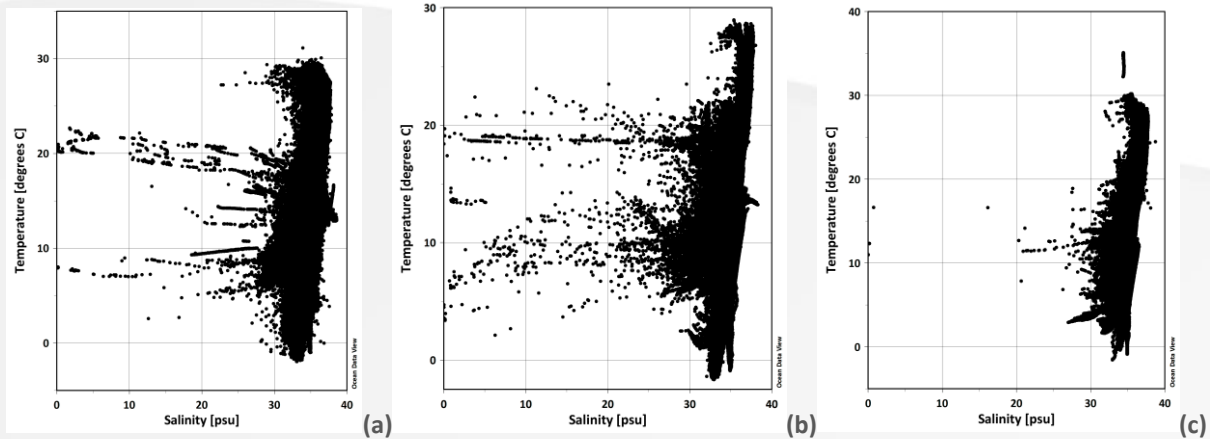


Figure 52 North Atlantic TS data collection excluding some bad data: (a) for the time period 1900-1999; (b) for the time period 2000-2008; (c) for the time period 2009-2014.

The procedure for the quality assessment analysis is built according to some criteria.

- Outliers: data outside of the regional range defined for temperature and salinity parameters are excluded (QF 4).
- Density inversion, when temperature and salinity measurements are available.
- A visual QF control is applied on all the dataset (with ODV), by time period, by latitude and longitude bands to focus on specific area. For all the spikes, density inversions, gradients, doubtful data, that are detected, the QC is changed to 3 (probably bad) or 4 (bad).
- Data on land have been identified.

The list of anomalies has to be sent to the NODCs.

9.6.2. The North Atlantic Restricted dataset

The new aggregation procedure considered also the restricted observations, whose distribution map is in Figure 65. For the North Atlantic area, there are not a lot of restricted data; they represent 2.63% of the whole aggregated data set.

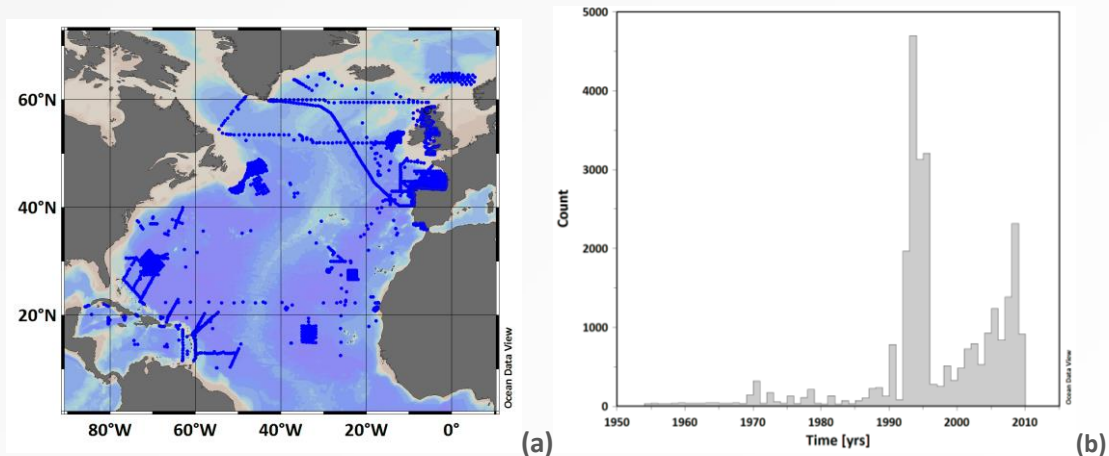


Figure 65 Restricted TS data collection for the North Atlantic Ocean in the time period 1900-2013: (a) Data distribution map; (b) Annual data distribution.

The quality of the restricted dataset is good except for few data; there are still bad data with QF1 and data with QF0 (Figure 66).

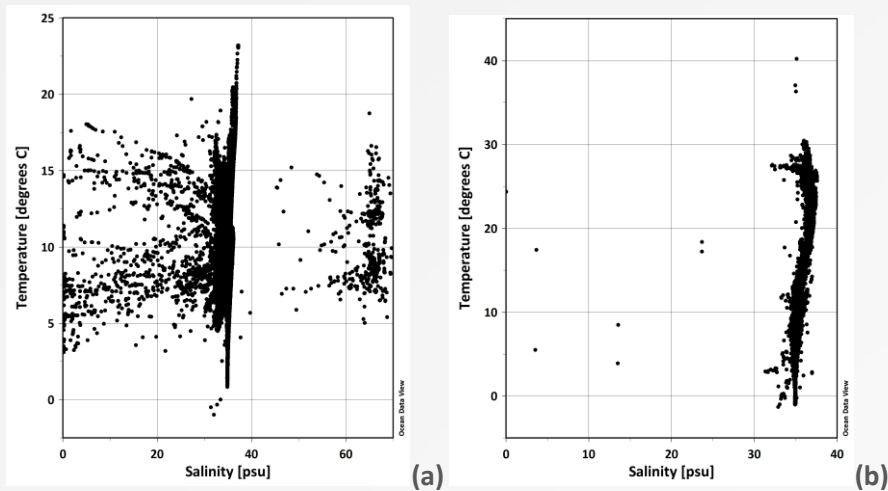


Figure 66 Restricted TS data collection for The North Atlantic Ocean in the time period 1900-2013: (a) TS diagram considering only data with QC flags = 1 (good) for T and S; (c) TS diagram considering only data with QC flags = 0 (no quality control) for T and S.

9.6.3. Conclusions

The quality of the data set from the North Atlantic Ocean is high. Less than 0.3% of all data were identified as suspicious after a detailed quality control. Most of the data are unrestricted, only a small portion (<3%) are restricted. List of anomalies has to be sent to each concerned NODC to correct their QF on data.

The density of the data is higher in the east part of the North Atlantic Ocean than in the west part that could introduce high error for the west part in the upcoming climatologies.

10. General Conclusions on the quality of the aggregated data sets

Aggregated datasets are available under

<ftp://ftp.ifremer.fr/ifremer/sismer/SeaDataNet2/Products/>

RCs analysed V1.1 regional data collections starting from the assessment of SDN data population increase with respect to V1. V1.1 data collections show a general data population increase due to the insertion of new data especially in the Atlantic region, as shown Tab. 11.

REGION	V1	V1.1	increase %
Atlantic	431974	1049547	143
Baltic Sea (#points)	8900000	11700000	31
Black Sea (1990-2013)	21068	23142	10
Mediterranean Sea	136828	169438	24
Arctic	407711	445281	9
North Sea			

Tab. 11 Number of data for V1 regional data collections, V1.1 data collections and the relative percentage of data increase.

The number of data in **North Atlantic area** has strongly increased between V1 and V1.1 versions (142%) and most of the new data come from the data centres listed in Table 2. Figure 67 shows an example of the new data coverage by Marine Institute (Ireland) between V1 and V1.1. However, due to the large increase of data, anomalies on QF are still observed on the data. Data with QF zero are still in the dataset.

Edmo_code	2013 (V1)	2014 (V1.1)	Data increase (%)
43 (BODC)	90213	331786	241573 (267.8%)
353 (IEO)	12345	188078	175733 (127.4%)
396 (MI)	3806	127592	123786 (3252.4%)

Tab. 12 List of EDMO codes of data centers that mostly provided data between V1 and V1.1 in the North Atlantic region.

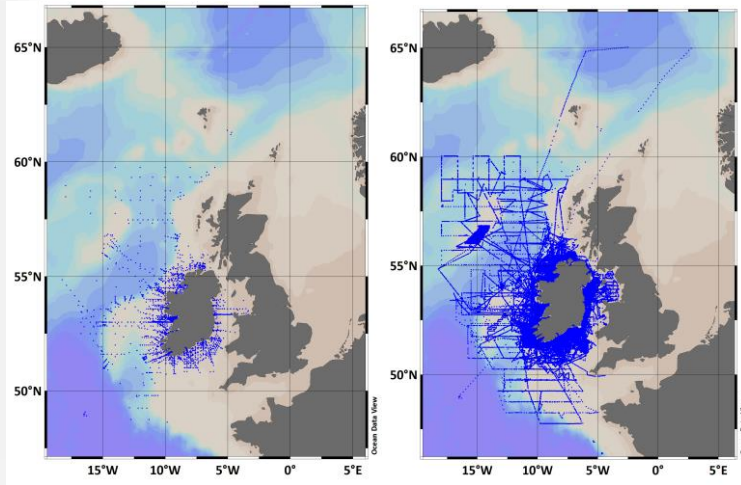


Figure 67 Example of the new data coverage by Marine Institute (Ireland) between V1 2013 (left) and V1.1 2014 (right).

Restricted data harvested during the 2nd aggregation procedure further increased V1.1 data collections with different percentages. In particular Black Sea and Mediterranean regions present still a conspicuous percentage of restricted data (19-17%), which are fundamental for climatology computation (see Tab. 13). Mediterranean restricted data, that are not included in the data harvesting process, are still a lot (estimate of another 20% from web portal), thus we will push data providers to allow internal access to those data for the Project purposes of statistical products computation.

	V1.1	Restricted	%
Atlantic	1049547	28318	3
Baltic Sea	11700000	65000	1
Black Sea (1990-2013)	23142	4287	19
Mediterranean Sea	169438	28690	17
Arctic	445281	245	0
North Sea			14%

Tab. 13 Number of stations in V1.1 regional data collections which contain only free access data and number of stations in the restricted data collections extracted from SDN data base during the second aggregation phase. The right column presents the percentage of the restricted data versus the total data.

The quality of the data collections improved due to the first QC implemented in coordination between WP10 RCs, MyO INSTAC and the NODCs, which applied corrections on the original data within the CDI. The second aggregation phase allowed retrieving new inserted data and the restricted access observations. This second QC phase highlighted that there are still some bad data flagged as good erroneously and also that there are still data not checked. RCs are preparing new QC reports to be sent to NODCs in order to analyse the remaining anomalies and get next year new V2 historical data collections with an improved quality, according with the schema shown in Figure 68.

The main conclusions on the first aggregated data sets might be summarized in the following point:

- All V1.1 data collections contain more data than V1 version
- Quality assessment highlighted that there are still anomalous data (flagged as good but not good) and data not checked (QF=0)
- Instructions on corrections of metadata and data and elimination of duplicates should be sent to data providers for applying as soon as possible otherwise V2 of will contain the same errors
- RCs agree on the need of a second feedback to data providers for V2
- RCs participated to WP activities
- A user guide has to be extracted/adapted from D10.2 before the external release
- Dissemination through SDN web catalogue (SEXTANT) requires some specific description.

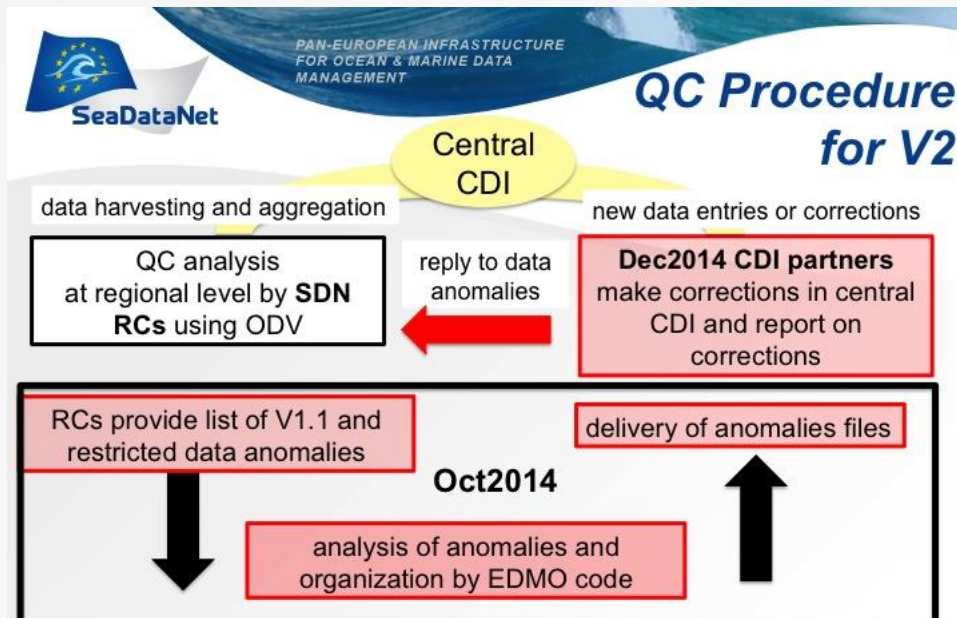


Figure 68 Quality Check procedure that is required before a new data aggregation procedure in order to provide the best SDN V2 final data collections.

References

S. Simoncelli, C. Coatanaon, Ö. Bäck, H. Sagen, S. Scory, D. Tezcan, D. M.A. Schaap, R. Schlitzer, S. Iona, M. Fichaut, M.Tonani. "TEMPERATURE AND SALINITY HISTORICAL DATA COLLECTIONS FOR THE EUROPEAN MARGINAL SEAS: AGGREGATION AND QUALITY ASSESSMENT PROCEDURES". IMDIS Conference 2013, Lucca, Italy.
http://imdis2013.seadatanet.org/content/download/73084/949811/file/S1P05_IMDIS2013_SDN2_Products_Simoncelli.pdf